Meeting of the Group of Experts on Business Registers
Online, 26 – 29 September 2022

Title: "Strategy and solutions for the EuroGroups Register (EGR)"
Authors: Merja Jalava, Enrica Morganti, Ioannis Sopranidis, Eurostat
Topic: 4. Modernization of the SBR

Group of Experts on Business Registers
Online meeting, 26 – 29th September 2022

Merja Jalava, Enrica Morganti, Ioannis Sopranidis

Eurostat

Session 4. Modernization of the SBR

Strategy and solutions for the EuroGroups Register (EGR)

Abstract

Globalisation creates continuous challenges to the production of economic statistics and ESS users requires a high quality and up-to-date EuroGroups Register (EGR) as the necessary infrastructure for the exchange of data on multinational enterprise (MNE) groups and for the coordination of globalisation statistics.

The EGR is the European statistical registers on MNE groups. 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.

The EGR is managed by Eurostat according to a well-established annual process since 2015. Its coverage and quality has improved significantly over the last years. The latest frames account for more than 140 000 multinational groups, comprising over 1 million legal units and more than 900 000 enterprises.

With limited resources available and growing requests to react fast to changes in statistics, it is necessary to review the quality priorities of the EGR. In order to guarantee better timeliness and higher accuracy for a limited set of top MNE groups with a sizable effect on European statistics, while ensuring a good coverage and accuracy (but less timeliness) for the rest, it was necessary to adapt the concept and the process of the current EGR. Therefore, the development of a new version of the EGR was launched, building on what is already achieved but also seizing the opportunity to embed new technical solutions and innovation.

The purpose of this paper is to outline the approach, objectives and solutions of these developments aiming at a new version of the EuroGroups Register. The proposal is based on an analysis of stakeholder's interests, feasibility and cost-benefit analysis. It implements also the elements of the ESS globalisation strategy related to statistical business registers, while upgrading the EGR system and process, including innovation, use of new data and new technologies and supporting a systematic approach to MNE groups data treatment in European Business Statistics (EBS).

Keywords: Multinational enterprise groups, Globalization, Statistical business registers

Introduction

The EuroGroups Register (EGR) is the statistical business register created by Eurostat to support the production of globalisation statistics in the EU. Eurostat receives input micro data from the National statistical institutes (NSIs) of European Union Member States (MS) and EFTA countries as well as from a commercial data provider and consolidates them into one European statistical business register on multinational enterprise (MNE) groups. The EGR data are accessible to statistical authorities in the EU and EFTA (i.e. national statistical offices) and national central banks of EU, European Central Bank, for statistical purposes.

The statistical output of the EGR are mainly twofold

1) Statistical frame of micro data, disseminated to all statistical authorities in the EU and EFTA (i.e. national statistical offices) also to EU national central banks, European Central Bank, used for the production of consistent statistics on globalisation;

2) Aggregated statistics on the structural characteristics of the multinational groups (by control, size, geographical presence, activities, etc.).

The data exchanges between the NSIs and Eurostat to produce the EGR are regulated by the European Business Statistics (EBS) Regulation 2019/2152[1]. The EGR is defined by the Regulation as the "authoritative source" (i.e. the sole provider of data records) for deriving high quality and harmonised populations for the production of European globalisation statistics.

The EGR is a well-established process since 2015, whose coverage and quality significantly improved over the last years. The latest disseminated frame accounted for about 140 000 multinational groups operating in the European Union and EFTA countries, comprising over 1 million legal units, more than 900 000 enterprises and 45 million employees. The EGR is a joint ESS product and its quality relies on several collaborative activities carried out in the ESS, e.g. data validation, European profiling, etc.

Users requesting access to EGR have been significantly growing over time and their requirements are becoming more and more ambitious thanks to the high potential of this data source in several statistical production processes. Thus, to continue supporting now and in the future the users with the highest quality, Eurostat launched the development of a new version of the EGR. The purpose of this paper is to outline the drivers, the strategy and the solutions for delivering this new version of EGR for the benefit of the whole ESS.

The drivers for change: who and why?

Globalisation is significantly affecting European economies and the phenomenon continuously generates new and timely information requirements from users and policy makers. The value added of EGR as core European register on MNE groups is highly recognised and over the years, more and more statistical producers became active EGR users in order to carry out their statistical work.

Statisticians in the ESS may need EGR for different purposes. The main and more general use of EGR remains as a survey frame for populations of units according to the "authoritative source" role assigned by the EBS Regulation. This role shall be fully implemented for the regular production of EBS dealing with globalisation. In addition, EGR is used for linking and checking information with other sources and as a coordination tool for analysing MNE groups' data consistency across domains. Finally, EGR is also used in combination with other statistical and administrative data to produce new output on globalisation by using micro data linking (MDL).

Having seen the importance of accurate and timely data for the largest and most significant MNE groups, many European statistical offices have created so-called Large Case Units (LCUs), or similar organisational arrangements to deal with data consistency. LCUs are horizontal units gathering experts with different specializations, responsible to systematically analyse limited sets of relevant MNE groups' and their data consistency across domains in a cost-effective way. LCUs need EGR information to deal with consistency issues that cannot be solved nationally. Since 2021, LCUs in the ESS are part of a "European Network of MNE groups

[1] Repealing 10 legal acts including Regulation 177/2008 on statistical business registers in the Union.

coordinators", i.e. a forum to share methodology and solve consistency issues having a sizable impact on data of different EU countries[2].

A more coordinated and systematic approach to MNE groups data in the ESS needs continuous improvements of the EGR which is at the core of the LCUs consistency work. In addition, in the area of business statistics, there is some clear intention to further reduce the deadlines of some statistics or provide early estimates, which require higher accuracy, more frequent updating and timely registers.

With limited resources available in the NSIs and at Eurostat, there is a need to focus efforts when trying to improve further the EGR quality and to prioritise some quality dimensions as not all the improvements can be carried at the same time and/or for the whole MNE groups' population.

The strategy for the new version of the EGR allows to upgrade the process and system and make it fit-for-purpose to the new requirements mainly re-using the existing capabilities and in the short-term run. As user requirements and technologies will inevitably continue to change in the future, the EGR will need to continue evolving in the longer term, in particular by embracing technologies changes and making use of new available data sources.

The strategy: what and when?

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

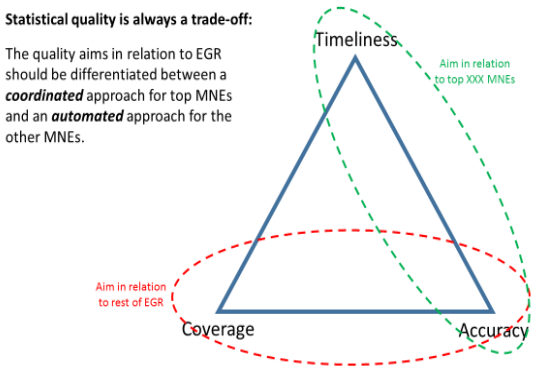
Therefore, with a good selection and prioritisation of the MNE groups according to their economic relevance, the EGR data quality management can be improved with a reasonable input of resources and a maximisation of results.

The strategy agreed by the European Statistical System is thus to ensure better timeliness and higher accuracy for a limited set of relevant MNE groups' which impact significantly the European statistics, while accepting only a good coverage and good level of accuracy for the smaller MNE groups and paying less attention to timeliness for those groups. This so-called "two-tier approach" means in practice the possibility to split the EGR population in a top-tier (Top MNE groups) and a bulk-tier (the rest of MNE groups) and adapt the system so as to implement a different data quality management process for the update of the two populations.

The top-tier population consists of the set of largest and most complex MNE groups with a significant impact on economic and business statistics, which need to be regularly monitored and frequently updated to provide users with high accuracy and a good and up-to-date statistical register infrastructure. The bulk-tier population includes all the other European MNE groups that will be updated only annually using mainly automated procedures at the best possible extent. This second set comprises the vast majority of data in the EGR. The bulk-tier is updated for the production of annual frame and provides a good coverage and sufficient accuracy to allow EGR serving as the authoritative population frame for the production of globalisation statistics, according to the EBS Regulation.

[2] The Early Warning System (EWS) is a structured, light (non-legislative) procedure, based on voluntary cooperation between national data compilers and Eurostat. It relies on a network of national EWS correspondents, coordinated by Eurostat EWS secretariat, who work on concrete restructuring cases.

Referring to the above-mentioned quality dimensions, the two-tier approach implies the prioritisation for the maintenance operations of the MNE groups in the EGR as sketched below.



The implementation of the two-tier approach is done by re-using to the maximum extent possible the available EGR IT system as the reference infrastructure, and adapting it to the new needs, adding new functionalities and taking the opportunity to introduce new data sources and new technologies. What is relevant is the distinction between the two ways of maintaining the MNE groups, based on their impact on statistics.

With more timely and better quality for the top-tier, the EGR frame can support the work of the LCUs and facilitate shortening the production time of some EBS. Good coverage and accuracy on the bulk-tier remains necessary to implement the EGR as the authoritative source as laid down in the regulation, and to allow using the EGR via micro data linking in order to produce new statistical output.

To discuss the technical modifications needed to be implemented in the whole EGR system (i.e. which thus includes the micro-data exchanges with the NSIs), Eurostat set up a dedicated Task Force in 2021, with some national experts. The Task Force worked on the methodological and technical solutions to implement in practice the two-tier approach described above.

Solutions: how?

The necessary modifications to the EGR system to implement the two-tier approach are already on going. The new version of the EGR will meet the new requirements with additional or improved functionalities already in the short run, without substantial changes to the existing infrastructure.

In particular, the identified solutions include several important building blocks.

Firstly, **a Complexity and Statistical Impact (CSI) index** has been defined and will be implemented in the EGR. This index will be calculated on all MNE groups present in the EGR, in order to distinguish the two populations (Top-tier and Bulk-tier).

The Complexity and Statistical Impact (CSI) index has been proposed by the Task Force and the criteria used for the CSI index are based on the experience already developed in some MS to select the MNE groups' populations that are treated by the Large Case Units.

The CSI index will select the top-tier MNE groups according to variables of the EGR and the result of the selection verified by the MS[3].

[3] The variables considered for the CSI are: the number of legal units, the number of different Economic Activities (NACE codes), the number of cross border relations (as number of countries with active units), the number of employees, the number of employees in largest legal unit, the number of levels of control in the structure of the MNE groups. For each variable, the MNE groups are divided into 10 equally sized groups (decile groups) and assigned a value between 1 and 10, 10 being assigned to MNE groups in the highest decile group. Employment is weighed higher by applying a value between 1 and 25 (in 25 equally sized groups). An overall value of the CSI index is then calculated for each MNE group present in EGR. An MNE group that is highest in the rank for all variables will have a total CSI index equal to 75.

Once the top-tier MNE groups are selected according to the CSI index, these MNE groups will be profiled and regularly monitored according to an event-driven model. The event-driven model has the objective to follow up the profiled top-tier MNE groups over time in a cost-effective way. According to the past experiences, the structure of MNE groups is quite stable and once they have been correctly profiled one year, it is usually not needed to profile them again the year after. The event-driven model shall indicate which MNE groups need to be updated by follow up profiling, if affected by events with impact on statistics. At the same time, it will allow to profile new top-tier MNE groups every year according to a growth model. With the introduction of an event driven model, the top-tier MNE groups will be followed up from one year to the next[4] and the EGR activities will not be necessarily sequential as they are now. Collection, process, validation and profiling could be done throughout the year, when the available information becomes available.

Secondly, **a leaner profiling process** has been proposed by the Task Force and it is currently still under discussion.

In fact, the current EU profiling, though a well-established collaborative activity done on a voluntary basis by EU MS and EFTA countries according to a 3-years programme and considered of high quality, is regarded by the NSIs as a too costly process in terms of resources. The results of EU profiling will be automatically integrated in EGR as part of the improvements. By end of 2023, about 300 largest and most significant MNE groups will be covered by the 3-years programme. However, according to the CSI index, about 2000 MNE groups shall be included into the Top-tier in the future.

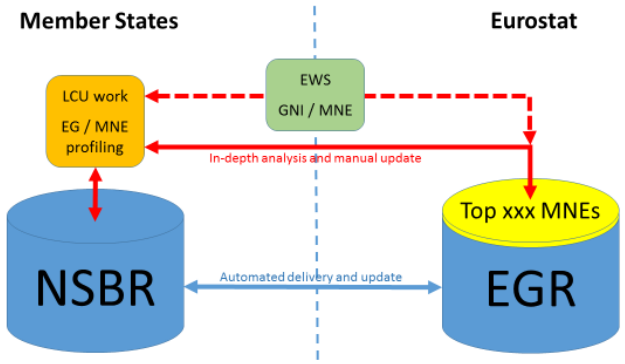
A leaner profiling process, compared to the current European profiling, is the alternative to increase gradually but significantly, the number of MNE groups in the top-tier. This lean profiling should aim at requiring less effort and resources and allowing a more cost-efficient updating process. This way the EU MS and EFTA countries will remain responsible for the defined number of top-tier MNE groups and will be able to update them according to the new lean process requirements. At the same time the existing European Profiling Programme shall remain in place and used according to the available resources and needs.

Finally, a **process automation for the treatment of the bulk-tier** will be implemented. Automation shall concern all data exchanges' operations between the national business registers and the EGR (in input and output), the implementation of SDMX (already available in 2022), data validation and quality checks, MNE groups' repairing phases, errors' reporting, etc. Automation aims at ensuring that the national registers smoothly integrate EGR data. Ultimately, it should liberate resources within the ESS so as to focus on the treatment of the top-Tier MNE groups.

Process automation will consider also the possibility to use new (structured and unstructured) data sources from open/public sources, to complete missing information in EGR and in particular increase the coverage for the extra-EU part of the MNE groups. As already mentioned, achieving a high coverage and good accuracy for the bulk-tier is an important pre-requisite in order to use the EGR as authoritative source for EBS and as tool to produce direct statistics by linking with other micro data. The possibility to produce new statistical output from the EGR could be enhanced using statistical modelling and other data science techniques opens up to new uses of the EGR in a cost-effective way.

The following picture describes at high level the interactions between the national and European registers and the different statistical activities.

[4] Technically this requires open up an EGR early initial frame immediately after the freezing of each EGR final frame, without waiting for the finalisation of the complete yearly process.



Research agenda and continuous improvements

As user requirements and technologies will continue to evolve over time, the EGR need to remain up to date and fit-for-purpose to serve them. Innovation shall also be included, in terms not only of technologies and data sources but also in term of organisation of processes to improve efficiency and deliver relevant output.

The research work to collect and use new data from publicly available sources for complementing missing information in EGR is ongoing. Web scraping techniques and new technologies will be embedded and support the production of EGR in a highly automated and timely manner.

The possibility of substantial evolutions of the EGR system is to be considered in a not-far future. In line with the ESS IT Strategy and innovation agenda, the EGR shall evolve into an infrastructure more and more based on new technologies and data sources. The intention to adopt modern digital solutions for statistical production, and to define a data architecture that leverages cloud platforms will open up in the next years the possibility for EGR to introduce new data processing and analytics technologies, and use modern platforms based on hybrid cloud technologies.

EGR will seize the opportunity to improve the processing, the visualization and the interaction with the data for users, and to provide tools for statistical analysis. In this context, the introduction and use of a modern data analytics platform as a visualization and presentation layer of EGR will be explored, allowing user interaction related to retrieving, visualising, and analysing the data. This will improve the way statisticians can access EGR data using modern dashboards, customized for the needs of each specific user group, without requiring extensive development for each specific data request. Data analysis tools will enable statisticians to analyse data directly, without needing to export them outside the platform.

At the same time, the enrichment of the existing EGR with machine learning capabilities, available in new platforms and tools will be investigated in order to improve the quality and completeness of the output. For the top tier, the implementation of a decision support system will be explored. This will allow users the possibility to complete missing data from the traditional EGR input sources, using machine-learning techniques and have the option to select, analyse and accept/refuse these data. For both the top tier and the bulk tier, systematic errors in the data shall be better identified and treated to further improve EGR quality. Machine-learning techniques offer the possibility to correct these data in an automated way and prevent the repetition of similar errors in the future.

References

1. A. Bikauskaite, A. Götzfried, Z. Völfinger, The EuroGroups Register in Statistika: Statistics and Economy Journal (2019 No. 99 (1), pages 69 - 76)

2. Eurostat, Statistical Business Register dedicated section

3. Eurostat, European business statistics methodological manual for statistical business registers (2021)