Data integration methods and tools in the ESBRs

Keywords: Globalization, interoperability, data integration, business registers, EuroGroups Register, profiling

1. INTRODUCTION

The **European system of interoperable Statistical Business Registers** (ESBRs) project is one of the ESS 2020 Vision Implementation Projects aimed at improving the quality of statistics in the EU.

In the ESBRs, Eurostat and the European Statistical System (ESS) partners cooperate by exchanging and integrating micro data on legal units, control relationships between legal units and enterprises to achieve a complete view on the structure and activities of multinational groups operating in the EU. The need to exchange statistical information on multinational groups comes from the fact that each national statistical office alone is unable to derive a complete and correct picture from its national administrative sources. They can observe only a 'truncated' view of the multinational groups for the legal units that are resident on their territory and some cross border relationships, while information about non-resident legal units and the control chain outside its territory is usually not reachable.

The integration process is managed centrally at Eurostat and takes place in the EuroGroups Register system (EGR 2.0 [1]). The national statistical business registers send their input data that cover EU legal units, while other commercial sources are used to cover non-EU legal units. In addition, national statisticians in different NSIs can improve the results automatically generated by the EGR. In particular, they further integrate statistical information obtained by manually profiling some of the largest and most relevant multinational groups in the EU. The result of European profiling is subsequently integrated into the EGR and constitutes another important pillar of the whole integration process in the ESBRs. This step of the ESBRs data integration process generates additional challenges because the view adopted in European profiling is top down, while the EGR integration process works bottom up.

The output of this integration process is a statistical frame, called the EGR global frame, containing the consolidated legal structure of the multinational groups in the EU and their statistical units. The EGR global frame is sent back to the national statistical institutes (NSIs). The feedback can be used at national level to be integrated back into the national statistical business registers to improve and complete their partial view on the multinational groups. Ultimately, the EGR global frame should function as the coordination tool for all ESS statisticians to improve the quality and consistency of data measuring the activities of the multinational groups across the EU.

2. METHODS: THE BUSINESS ARCHITECTURE APPROACH

Any data integration project presents many challenges and requires that some aspects are carefully taken into account to ensure the quality of the output. Data coming from different sources are usually not harmonised in terms of concepts, definitions and data

structures and require methodological analysis and transformations before they can be used for integration.

Despite the ESBRs project built upon the existing EGR 1.0 data exchanges, EU Regulations and European Recommendation Manual, still the level of interoperability among the national statistical business registers and the European register was very low when the project was launched and negatively affected the quality of the output.

In order to take into account all the interoperability aspects and develop solutions that could increase the quality of the output, the ESBRs project adopted a holistic approach and created a ESBRs business architecture document describing the transition from the initial situation to the improved ones using European standard frameworks (EIF, EIRA [2]). The ESBRs can thus be viewed as an ecosystem where the information is continuously improved at all network nodes via continuous data exchange and data integration.

This paper will focus mainly on some changes introduced by the ESBRs business architecture and related to data integration aspects that worked in the direction of improving the quality of the output:

- 'Authentic source' principle. It refers to the fact that the ESBRs information is stored only once in the system and can be changed only by one role.
- 'Unique identification of information objects'. This aspect is one of the most crucial issues in integrating overlapping data representing the same real- entity, which requires a true identification. This implies the detection and elimination of duplicate records. In the ESBRs, all information objects have a unique identifier over time (the LEID number) that is assigned as the result of linking and matching algorithms. Human validation is also part of the identification process to ensure quality.
- 'Single flow' principle. This ensures that in the integration process all changes happening in data over time are handled consistently and in the same way.
- 'Interoperability'. It means that data has to respect a certain level of harmonization and standardization when exchanged between parties in order that they understand it in the same way. It is about preserving data meaning and it has to apply at conceptual, definitional and technical level.

Last but not least

- 'Workable' principle. Data integration in the ESBRs is limited to what is necessary for achieving the agreed objectives in respect of the principles of subsidiarity and proportionality of the European Member States and EFTA countries that provide their micro data as input to the process.

The paper will also briefly describe the EGR 2.0 'Priority rules' governing the integration process and the way one source is chosen among the many competing ones for the same record, unit or variable.

The ESBRs makes use of several tools and applications that are necessary for the data integration and/or support it with some specific functions (identification of legal units and LEID assignment, validation of preliminary results of data integration, European profiling to improve the quality of the output). All such tools and applications follow the principles mentioned in the bullet points above. They are described the second part of the paper.

3. ESBRs Tools – practical aspects on data integration

The ESBRs tools are a set of software applications developed under the ESBRs project:

- the EGR Identification Service (EGR IS): Application supporting the EGR users in identifying legal units.
- the EGR Interactive Module (EGR IM): Interactive web interface to browse and validate data in the consolidation area of EGR CORE.
- the EGR Foreign Affiliates interface (EGR FATS): Application providing users with a FATS statistics oriented web interface to browse and download EGR data.
- the EGR CORE application: The 'heart' of EGR 2.0 system, it stores, transforms and consolidates input data of different sources and generates the EGR frames.
- the Interactive Profiling Tool (IPT) prototype: Application supporting the users in their European Profiling[3] activities concerning large MNEs.

4. The ESBRs Interoperability pilots – testing of the methods and tools

4.1. Objectives

The objective of the interoperability pilots is to test elements of interoperability and eventually create best practices. These best practices are expected to lead to further integration of the network of SBRs and therefore strengthen their backbone role.

4.2. Areas of interest

Four areas of interest have been identified for the interoperability pilots:

A. Data exchange: These pilots focus on standardization and automation in the data exchange process between EGR/IPT and the national SBRs.

B. Collaboration and governance: These pilots focus on cross-border collaboration using common interactive tools, chiefly EGR and IPT.

C. Timing/calendar: These pilots focus on adaptability of national and European processes to yearly cycles for national SBRs, EGR, national Profiling and European Profiling/IPT and the interaction between these processes.

D. Quality: These pilots focus on output data assessment and input data improvement.

4.3. Supported ESBRs deliverables

The pilots employ the applications developed by the ESBRs project and comply with the ESBRs methodological deliverables, including the ESBRs Business Architecture (BA), the ESBRs Interoperability Framework (IF) and the European Profiling methodology.

4.4. IT applications and infrastructure

Relevant infrastructure and deliverables of other projects that can be reused during the pilots include:

- the VIP.ESDEN project and the relevant work on data exchange (TESTA network);
- SDMX tools: SDMX Converter, SDMX Reference Infrastructure (SDMX-RI).

5. CONCLUSIONS

The ESBRS project aims at improving the consistency, efficiency and interoperability of the data exchanges and data integration process between the national SBR and the central EGR register. Within this project, the business architecture approach designed the principles to improve the interoperability and the development of the tools that could improve the data integration process and ultimately the quality of the statistical output.

REFERENCES

[1] A. Götzfried, Z. Völfinger and A. Bikauskaite, "The EuroGroups Register", IAOS-OECD 2018 Conference

[2] EIF, European Interoperability Framework: https://ec.europa.eu/isa2/eif_en

EIRA, European Interoperability Reference Architecture: https://joinup.ec.europa.eu/collection/european-interoperability-reference-architecture-eira

[3] I. Xirouchakis and V. Hecquet, "Improving the quality of Business Statistics through Profiling", European Conference on Quality in Official Statistics, Krakow Poland, June 2018.