# Micro data linking

2019 edition

MANUALS AND GUIDELINES

eurostat

# Micro data linking

## 2019 edition

Manuscript completed in November 2019

*Printed by the Publications Office in Luxembourg*



Luxembourg: Publications Office of the European Union, 2019



PDF: ISBN 978-92-76-13178-6 ISSN 2315-0815 doi: 10.2785/568055 KS-GQ-19-013-EN-N

# Executive summary:

- Microdata linking (MDL) provides an opportunity to develop new statistics and indicators and to discover new information both when using existing data sets but also when combining them with new data collections. Together with Member States, Eurostat has been carrying out a number of microdata linking projects aiming at establishing methodologies and developing country practices for microdata linking. The results of the efforts have been very rewarding in terms of new statistical indicators that could produce additional insight into the performance of enterprises.
- These guidelines provide an overview of past experiences and best practices for the implementation of microdata linking methods. The starting point is the most recent exercise of linking Business demography (BD) and Trade by enterprise characteristics (TEC) data. In addition, this exercise made use of Structural business statistics (SBS) as the central data source for microdata linking projects concerning enterprises.
- This publication introduces the rationale and key elements of microdata linking and suggests new angles for economic analysis and support for policy making. It also deals with the benefits and limitations of microdata linking and the main lessons learned so far.
- The key concepts and sources are described in broad terms. The central role of the Business register is emphasized when carrying out microdata linking exercises. Data linking with trade statistics including the possibilities of using either Trade by enterprise characteristics (TEC) data or International trade in goods statistics (ITGS) as well as Business demography (BD) and Statistics on foreign affiliates (FATS) is presented in this publication.
- Several aspects related internal and external data validation are discussed as well as methods needed to impute missing information.
- In order to finalize the outputs of MDL, considerations on confidentiality and revisions are discussed as a part of the dissemination strategy.
- In the future, combining the enterprise data to social statistics through variables such as educational attainment, gender, income/salaries, years of work experience and more is expected to enrich and deepen linked microdata analysis. This would allow new kind of analysis, linking business dynamics to social aspects, such as gender or income inequality.
- Microdata linking provides unique opportunities to create more insight into the business economy without increasing the burden on respondents. This means that microdata linking will become increasingly relevant in the future.

# Contents

# Preface

These guidelines collect together experiences of the recent micro data linking exercise of linking Business Demography (BD), Structural Business Statistics (SBS) data together with Trade by Enterprise Characteristics (TEC) data. The purpose of these guidelines is to serve as a first guidance to the theme of micro data linking i.e. what kind of challenges are expected and which kind of solutions could be applied. These guidelines are a useful tool for the countries with no prior experience of micro data linking. Eurostat facilitated the project involving the following national statistical institutes listed below and the OECD. Eurostat appreciates the contributions of all participants.

Danmarks Statistik – Statistics Denmark
Statistikaamet– Statistics Estonia
Centraal Bureau voor de Statistiek – Statistics Netherlands
Instituto Nacional de Estatistica – Statistics Portugal
Tilastokeskus - Statistics Finland
Statistiska centralbyrån - Statistics Sweden
Statistisk sentralbyrå - Statistics Norway

# Glossary of abbreviations

| | |
|---|---|
| BD | Business Demography |
| CIS | Community Innovation Survey |
| EGR | EuroGroup Register |
| ESS | European Statistical System |
| EU | European Union |
| FATS | Foreign Affiliate Statistics |
| FRIBS | Framework regulation integrating business statistics |
| FTE | Full Time Equivalent |
| IBAS | Integrated Business Account System |
| ICT | Information and Communication Technologies |
| ID | Identification |
| IFATS | Inward Foreign Affiliate Statistics |
| ITGS | International trade in goods statistics |
| MDL | Micro Data Linking |
| NACE | Nomenclature of Economic Activities |
| OFATS | Outward Foreign Affiliate Statistics |
| PRODCOM | Statistics by Product |
| R&D | Research and Development |
| SBR | Statistical Business Register |
| SBS | Structural Business Statistics |
| SME | Small and medium sized enterprise |
| STEC | Services Trade by Enterprise Characteristics |
| TEC | Trade by Enterprise Characteristics |
| UCI | Ultimate Controlling Institutional unit |
| VAT | Value Added Tax |

# 1 Introduction

## 1.1. BACKGROUND

European business statistics compilers often face a dilemma: On the one hand, users and policy makers demand additional information on the structure and development of European enterprises. On the other hand, budget constraints and reluctance to increase the burden on survey respondents and national statistical institutes put tight restraints on the extension of data collection.

Microdata linking (MDL) provides an opportunity to discover new information and to develop new statistics and indicators both when using existing data sets but also when combining with new data collections. During the recent past Eurostat together with Member States have been carrying out a number of microdata linking projects aiming at establishing methodologies and country practices for microdata linking. The rationale has been to provide new statistical indicators based on linked data, and the results of linking efforts have been very rewarding. For instance, linking trade statistics with structural business statistics has allowed to investigate a suite of questions on the type of firms that trade, in particular their size and sector, bringing granularity in the analysis and monitoring of the performance of the business sector. The same information could not be provided by trade statistics alone or structural business statistics alone; due to the relevance of the insights it generates, the dataset Trade by Enterprise Characteristics (TEC) has become a key tool for analysis and policy, for example, on the small and medium-enterprises (SME) sector.

In light of its potential, microdata linking has been one of the priority areas in the business statistics in the recent past. While still limited, the number of EU countries involved in microdata linking is foreseen to be gradually increasing. Eventually the hope is to produce interesting data on European economies on a regular basis, reflecting increasingly complex policy needs.

Ideally microdata linked projects do not implement changes in data collections in participating countries, but build up a coherent plan on how to use microdata linking to create relevant datasets for further economic analysis. This also means that participants in microdata linked projects have an output oriented mind and view on the most pressing policy questions to be reflected in microdata linking exercises.

These guidelines gather past experiences and best practices of the microdata linking, focusing on the most recent exercise of linking Business demography (BD) and Trade by enterprise characteristics (TEC) data. In addition, the guidelines highlights Structural business statistics (SBS) as the central data source for microdata linking projects concerning enterprises.

There are several benefits from microdata linking, however also some limitations:

**Benefits:**

Microdata linking is an effective tool for compiling new statistical data. Firstly, it has significant potential in gathering new statistical evidence without increasing the burden placed on respondents. Respondent burden is a major policy issue needing constant attention by national statistical authorities. Indeed microdata linking can reduce the burden on enterprises when conducting new or existing surveys by eliminating questions that can be answered by combining existing data from different sources.

Secondly, microdata linking can be used for ensuring data quality and consistency between related data sets and as a tool to reduce discrepancies that occur in mirroring international statistics e.g. in inward and outward foreign affiliate statistics.

A coordinated approach to microdata linking across the EU Member States is a cost-efficient way for national statistical offices to undertake microdata linking and ensure harmonised, comparable results across countries.

The challenging theme of globalization can be approached by using microdata linking techniques. This relates directly to the multinational enterprises and their global patterns to organize business functions. For example, microdata linking approach can provide evidence on global business activities, by generating enterprise statistics data which show, if the enterprise belongs to an enterprise group or not, and whether the enterprise group itself is all national or foreign.

**Limitations:**

It is important to decide on concrete objectives before launching a microdata linking project and to design the linked data set carefully i.e. decide on the variables for the analysis and the aggregates to be produced. This is often challenging since these are often pilot projects in nature and it is difficult to define project in details from the outset.

From the analytical point of view it may be desirable not only to link the latest data but also earlier records going back as many years as possible. Since Business demography events like the creation, termination or acquisition of enterprises affect the business registration population over time, the matching rates are the highest for latest data and will typically decrease the further we go back in time.

It would be important to have the same statistical units across countries. The choice of legal units versus enterprise versus enterprise group may complicate MDL, and hamper the comparability of the results. In the Framework regulation integrated business statistics (FRIBS), one common statistical unit, the enterprise, is foreseen to be implemented in business statistics.

*Note 1 – Benefits and limitations to microdata linking*

## 1.2. LESSONS LEARNED

Since several coordinated microdata linking projects have been carried out within the European Statistical System (ESS), it is appropriate to list the main lessons learned that are helpful for future exercises.

Organisational aspects and coordination plays an important role in microdata linking. This is because a coordinated microdata linking exercise implies that multiple countries are carrying out innovative activities at the same time, and try to achieve harmonized and comparable output from, in many ways, separate projects. This requires a great deal of knowledge of country specific issues going from IT systems to specific revision policies of the source data, or to organisational issues. For instance, the different data sources can be found in different administrative sections or units within National Statistical Institutes (NSI), or even in a different offices entirely (for example in Finland, TEC is produced by the customs office). This is why it is recommended to have highly skilled national experts involved that can efficiently operate in the NSI and collaborate with the international partners.

Three possible approaches to microdata linking:

- Fully coordinated approach is an efficient and feasible way to produce micro linked data, which means that each country runs harmonized statistical software codes produced centrally with commonly agreed definitions and breakdowns. However, this requires compatible IT systems and identical databases. Currently this is easily achievable only in the Nordic countries.

- Another approach to achieve similar results, although with increased uncertainty over comparability, is to develop a set of instructions (or pseudo-code) that could be further developed by national experts to fit the IT environment in place.

- The third approach is to commonly agree on the set of definitions and breakdowns, and let each NSI develop codes to produce the output as requested. In this third case, more effort is needed to ensure harmonization by other means of quality control.

Registers where administrative data are the main source, with common identifiers across statistical domains, are more suited than sample based data sets for applying the microdata linking approach, because of the representativeness issues. However, it has been shown that the sample based systems can accommodate microdata linking approach successfully. In that case, an extra effort on extrapolation to the population totals is necessary and certain skill in methodology is required. See the examples from the Netherlands on this point in chapter 4.

A successful microdata linking project can be divided into three phases (see Annex 2).

- Firstly, matching and adjustment of data and structuring of database, which includes defining data sources, variables and reference periods.

- Secondly, validation controls and calculation of weights for the control group population(s) where necessary, which include, for instance, inspecting demographic events and relations and weights for control group according to the composition of the target population (which will be defined for a given analysis).

- Thirdly, production of standardized output, including descriptive statistics, longitudinal or regression analysis. Ideally, a policy relevant written analysis is produced to support the use of data.

## 1.3. OVERVIEW OF THE CHAPTERS

Chapter 1 introduces the rationale and key elements of microdata linking. In addition, it suggests new angles for economic analysis and support for policy making. It presents the benefits and limitations of microdata linking and the main lessons.

Chapter 2 provides an overview of the statistical unit and sources relevant for compilation of micro linked data. The key concepts and sources are described in broad terms. Potential sources for micro linked data covers Business register, Structural business statistics, Business demography and International trade statistics with additional references for other possible sources.

Chapter 3 emphasizes the central role of the Business register when carrying out microdata linking exercises. It describes in fairly detailed level the data linking with trade statistics including the possibilities of using either Trade by enterprise characteristics (TEC) data or

International trade in goods statistics (ITGS). Chapter 3 also covers the specific features of linking to Business demography (BD) and Statistics on foreign affiliates (FATS).

Chapter 4 deals with several aspects related to outputs of the project including internal and external data validation as well as issues related to imputation. Since the practices may vary considerably across countries, some boxes including country examples to impute or validate data are attached.

Chapter 5 on dissemination strategy, confidentiality and revisions introduces the key breakdowns used in the most recent microdata linking exercise. The chapter also introduces an example of output tables.

Chapter 6 concludes by presenting some closing thoughts about the microdata linking projects already carried out in the EES. It also reflects the possible future areas where the microdata linking methodology could be used and further elaborated.

# 2 Statistical units and sources

The statistical units for the compilation of linked microdata are derived from national statistical business registers (SBR). The statistical units in microdata linking are enterprise and enterprise group, whose definitions are based on an EU regulation concerning statistical units (EEC 696/93) and regulation on Business Registers (EC 177/2008). The adoption of an enterprise is in different phases across the European Statistical System (ESS). As a result earlier microdata linked statistics were typically compiled using the legal unit – in some countries this is still the case pending implementation of the enterprise as statistical unit. These guidelines assume that countries have adopted the enterprise as a statistical unit in Structural Business Statistics (SBS), and the registers are maintained at the enterprise level.

## 2.1 UNITS

An enterprise is an organizational unit producing goods or services which has a certain degree of autonomy in decision-making. An enterprise can carry out more than one economic activity and it can be situated at more than one location. An enterprise may consist out of one or more legal units.

Legal units include legal persons whose existence is recognized by law independently of the individuals or institutions which may own them or are members of them, such as general partnerships, private limited partnerships, limited liability companies, incorporated companies etc. Legal units as well include natural persons who are engaged in an economic activity in their own right, such as the owner and operator of a shop or a garage, a lawyer or a self-employed handicrafts-man.

The legal unit, and not the enterprise, is liable to report statistics to authorities. This is why, typically, the enterprise level information needs to be derived from the legal units and/or the group head.

A legal unit may own a second legal unit and this second legal unit may carry out activities solely for this first legal unit. E.g. legal unit A, a limited liability, produces particular goods and legal unit B, a limited liability as well, solely sells these goods. Both units have the same management. In this case they are seen as one single enterprise. Another example may be that legal unit C employs the staff and legal unit D owns the means of production like machines and buildings. A third legal unit E may own and manage these two legal units. Only the units C, D and E together can produce something and hence are to be counted as one enterprise.

Reasons for splitting the organisational unit enterprise into more than one legal unit can be manifold: optimizing taxes or liabilities, different salaries according to the collective wage agreement or avoiding the publication of annual reports are among them. For example, an enterprise with a certain activity might be able to save expenses, if the wages in the collective

wage agreement of this activity are higher than e.g. in logistics. In that case it could make sense to single out the transport capacities of this enterprise into a special limited liability. While in the practical organization of the enterprise nothing has changed, it now legally consists of two legal units.

An enterprise group is an association of enterprises bound together by legal and/or financial links and controlled by the group head. A group of enterprises can have more than one decision-making center, especially for policy decisions on production, sales and profits. It may centralize certain aspects of financial management and taxation. It constitutes an economic entity which is empowered to make choices, particularly concerning the units which it comprises.

The group head is a parent legal unit which is not controlled either directly or indirectly by any other legal unit. The subsidiaries are considered to be subsidiaries of the parent enterprise.

A multinational enterprise group is defined as an enterprise group composed of at least two enterprises or legal units located in different countries.

## 2.2 SOURCES

### 2.2.1 STATISTICAL BUSINESS REGISTER

Statistical business registers (SBR) plays a central role in the production of business statistics. Both in terms of the way the statistics are produced and in terms of the content and quality of the statistics. The availability of statistical business registers is essential to the compilation of consistent and comparable business statistics.

Statistical Business Registers are essential for establishing efficient statistical survey frames. A statistical business register makes the national statistical system more efficient and helps to reduce the reporting burden on businesses.

Most statistical business registers contain information on the population of the following statistical units:

- Enterprises performing economic activities that contribute to gross domestic product (GDP)
- Their local units
- The legal units that make up such enterprises
- Enterprise groups.

**The characteristics of the units that are recorded in the registers include:**

- Identification characteristics: ID numbers, names, addresses
- Demographic characteristics: date of commencement / cessation of the unit's activity
- Economic/stratification characteristics: economic activity (NACE), employment, turnover, institutional sector code, legal form
- Information on control and ownership relations: parent/subsidiary legal unit, minority shareholder information, country of global decision center.

*Note 2 – Basic information found in the Statistical Business Register.*

Additionally, many business registers hold information on establishments, which can be linked to enterprises. By linking the location information, it is possible to create statistics on the spatial dimensions of enterprise characteristics.

### 2.2.2 STRUCTURAL BUSINESS STATISTICS

Structural business statistics (SBS) can be used to address various questions related to following main questions:

- How much value added is created in a given sector and how many persons are employed.
- Describing the structural change in the business sector, following the number of person employed in a specific sector over time.
- Making comparisons of the average wages over time and between different sectors in the economy.
- How productive is a particular sector in an economy.

*Note 3 – Information provided by Structural Business Statistics.*

SBS covers business economy including industry, construction, distributive trades and services. Presented according to the NACE activity classification, they describe the structure, conduct and performance of businesses across the European Union. These statistics can be broken down to a very detailed sectoral level – several hundred economic activities. A subset of the SBS information is also available for European regions, as well as according to the size of enterprises. The main indicators within SBS are generally collected and presented as monetary values, or as counts (for example, numbers of enterprises or persons employed); this is in contrast to short-term business statistics, where the data are presented as indices – generally in relation to a base year. Generally SBS does not collect information on products. The external trade and the production of specific products are covered by external trade statistics and PRODCOM.

SBS may be broken down by different regions or by enterprise size-class. In SBS, size-classes are defined by the number of persons employed. A limited set of the standard SBS variables for example, the number of enterprises, turnover, persons employed and value added is available, mostly down to the three-digit group level of the NACE classification.

### 2.2.3 BUSINESS DEMOGRAPHY STATISTICS

In business demography (BD) statistics one of the main principles is to follow the organic development of enterprises (a cohort), over time. In general this means that administrative entries and exits, that usually bias the real dynamics of enterprises, are excluded. In microdata linking tables, the BD is used as a source for defining the enterprise age, thus allowing to carry out analysis with an age variable that is, for example, clean from non-real entries resulting from administrative restructuring, where already existing resources are simply transferred between enterprises.

BD is a qualified extraction from the SBR: It does not purely follow the SBR or administrative sources, but the BD rules are applied. In the EU, the Eurostat-OECD BD manual is used as a basic guidance to compilation of Business Demography Statistics.

Business demography statistics present data on:

- The active population of enterprises
- Their birth
- Survival (followed up to five years after birth)
- Death
- Enterprise growth

**Special attention is paid to the impact of these demographic events on employment levels. Business demography data can be used to analyse the dynamics and innovation of different markets, for example:**

- Entrepreneurship in terms of the propensity to start a new business
- The contribution of newly-born enterprises to the creation of jobs
- Number of high growth enterprises (and employment levels)
- Number of young high growth enterprises

*Note 4 – Information provided by Business Demography Statistics.*

Business demography delivers key information for policy decision-making and for the indicators to support the Europe 2020 strategy. It also provides key data for the joint OECD-Eurostat "Entrepreneurship Indicators Programme". Enterprise birth rates, death rates and two-year survival rates form part of the structural indicators. The concepts and definitions as well as guidelines on how to implement the statistics are explained in detail in the Eurostat-OECD Manual.

### 2.2.4 INTERNATIONAL TRADE STATISTICS

International trade statistics are available in many data sources. Currently, the main statistics concerning exports and imports are International Trade in Goods Statistics (ITGS), Trade by Enterprises Characteristics (TEC), International Trade in Services Statistics (ITSS), and Services Trade by Enterprise Characteristics (STEC).

International trade in goods statistics (ITGS) measure the value and quantity of goods traded between EU Member States (intra-EU trade) and goods traded by EU Member States with non-EU countries (extra-EU trade) break down by types of goods (Combined Nomenclature) and by partner countries. "Goods" mean all movable property, including electricity and natural gas.

The providers of statistical information differ between intra and extra EU-trade. In the first case, it corresponds to all taxable persons reporting transactions exceeding a certain threshold fixed by member states; in the second one, it corresponds to administrative data from the customs declarations lodged by natural or legal persons in the customs administration. Eurostat publishes aggregated and detailed data for ITGS: The former comprise monthly and annual data with a breakdown of large product categories and by partner country, the latter refer to monthly trade statistics at the most detailed level of several products and also by partner country.

The methodology of ITGS is explained in detail in the "**Compilers guide on European statistics on international trade in goods".**

The main objective of the trade in goods statistics by enterprise characteristics (TEC) is to bridge two major statistical domains which have traditionally been compiled and used separately, business statistics and ITGS. Specifically, this new domain was created to answer questions such as:

- What kind of businesses are behind the trade flows of goods?
- What is the contribution of a particular activity sector to trade?
- What is the share of small and medium-sized enterprises to total trade?

*Note 5 – Information provided by Trade by Enterprises characteristics (TEC).*

For this purpose, the trade in goods between countries is broken down by economic activity, size-class of enterprises, trade concentration, geographical diversification and products traded. The new information is used to carry out more sophisticated kinds of analysis, e.g. to evaluate the role of European companies in the context of globalisation or to assess the impact of international trade in goods on employment, production and value added, essential in a globalised world where economies are increasingly interconnected.

International Trade in Services Statistics (ITSS) currently covers the monetary indicators for trade in services broken down by the service categories (e.g. computer services, legal services etc.) and by partner countries. These statistics come from the transactions recorded under the country's balance of payment, hence the transactions that take place between economy's residents and non-residents. The current microdata linking project covers only the trade of goods and excludes trade of services.

Services Trade by Enterprise Characteristics (STEC) is a microdata linking exercise in itself, with the aim of linking characteristics of the services traders to SBR and SBS, providing important insight on what kind of enterprises are behind the service trade. In other words, what kind of actors are there, and possibly, what kind of policy tools could be relevant in targeting the services traders and service producers of the economy. STEC has provided statistical output describing for example, the roles played by Small and Medium sized Enterprises (SMEs) and large enterprises alongside their industry and ownership (foreign or domestic).

### 2.2.5 FOREIGN AFFILIATE STATISTICS (FATS)

The legal basis for FATS in the EU is the European Parliament and the Council Regulation (EC) No 716/2007. FATS is distinguished into “inward foreign affiliate statistics” and “outward foreign affiliates statistics”. The former describe the activity of foreign affiliates resident in the compiling country, while the latter describe the activity of foreign affiliates abroad controlled by the compiling country.

Inward FATS aim to assess the impact of foreign-controlled enterprises on the European economy and in particular to measure the impact of foreign control on employment, wages and productivity. Outward FATS measure the commercial presence through affiliates in foreign markets. In particular, outward FATS data measure the turnover, activity, number of persons employed and number of foreign affiliate controlled by the EU member states. The FATS are useful for answering following questions:

- Is this enterprise domestically owned or foreign owned?
- From which country is this enterprise owned?
- How many employees does this enterprise have abroad?

*Note 6 – Information provided by FATS.*

### 2.2.6 OTHER STATISTICS

The microdata linking projects have considered other statistical domains to produce experimental output, and there exist other sources, which can be interesting extensions for the future.

The Community Innovation Survey (CIS), the basis for innovation statistics, is part of the EU science and technology statistics. Surveys are carried out with two years' frequency by EU member states and number of ESS member countries. Compiling CIS data is voluntary to the countries, which means that in different surveys years different countries are involved.

The CIS is a survey of innovation activity in enterprises. The harmonised survey is designed to provide information on the innovativeness of sectors by type of enterprises, on the different types of innovation and on various aspects of the development of an innovation, such as the objectives, the sources of information, the public funding, the innovation expenditures etc. The CIS provides statistics broken down by countries, type of innovators, economic activities and size classes. The survey will also be part of the future data collection under FRIBS.

Research and experimental development, abbreviated as R&D, refers to creative work undertaken on a systematic basis in order to increase the stock of knowledge (including knowledge of man, culture and society), and the use of this knowledge to devise new applications. The data are obtained through statistical surveys, which are regularly conducted at national level covering R&D performing entities in the private and public sectors.

The Community survey on ICT usage and e-commerce in enterprises is an annual survey conducted since 2002, collecting data on the use of information and communication technology, the internet, e-government, e-business and e-commerce in enterprises. The survey will also be a part of the future data collection under FRIBS.

Many countries have compiled linked employer -employee data for research use. Typically, the data include working populations' basic characteristics, such as family, living, employment relationships, periods of unemployment, income and education. Based on the employment relationship at the end of the year, the data on personnel are linked to various data files on enterprises, e.g. the financial statement panel. Examples of these data are the Finnish Longitudinal Employer-Employee Data (FLEED) the German matched employer–employee dataset (LIAB), Norwegian Linked Employer-Employee Database (LEED) and the Danish Matched Employer-Employee Data. These types of linked data sets could provide a basis for very interesting extensions using the microdata linking methodology.

# 3 Enterprises and linking to SBR

This chapter deals with several aspects of microdata linking, all focused on the central position of the SBR in linking data from various sources, including other statistical registers and administrative sources. A brief review is given of methods for dealing with combining incomplete datasets and ensuring consistency with published data. Some country practices are described in separate boxes.

## 3.1. THE CENTRAL ROLE OF THE SBR

Statistical business registers cover a limited set of variables on businesses, mostly limited to direct use relevant for the business register and survey samples taken from it. Key variables for a statistical business register are economic activity - determined either as the presence of employees or turnover – and a classification of this activity and size measured in persons employed. Demographic events such as mergers/acquisitions, births and deaths can also be obtained from the SBR.

Linking these data from the Statistical Business Register on micro level, to a wide variety of administrative data, other data sets, or survey data can provide new insights into the relationship between enterprise characteristics and enterprise performance.

The SBR is at a central position in the statistical process: At the start it provides survey frames. At the end of the statistical process the output consists of a selection of the SBR population with values assigned to the relevant survey variables.

Information in the SBR is available at several hierarchical levels: the Enterprise Group, Enterprise, Fiscal Unit, and Legal unit. The recommended statistical unit is the enterprise, which means that sometimes legal units need to be grossed up in order to represent the enterprise. Many variables may not be directly additive, and different country practices exist in order to handle this issue. In most of the countries participating in microdata linking, the legal unit has been used directly as a proxy for an enterprise.

In the Netherlands the input for the Business Register are legal units and the ownerships relations between them. These ownerships relations are obtained from the fiscal authorities. VAT units are allowed to form VAT groups - to prevent unnecessary taxation of inter-company transactions - when certain conditions are met: financial, economic and organizational intertwining of the VAT units. This is relevant for ITGS, as the VAT data is collected at the level of the VAT entity (either a single VAT unit or a VAT group). VAT groups and VAT units are combined into enterprises based on ownerships relationships of >50% in the underlying units to align with the concept of statistical unit.

*Note 7 – The hierarchy of the Dutch SBR*

Using legal units or enterprise has no significant consequences for these guidelines, as the process of linking is not affected, as long as the unique identifier is available across the registers.

## 3.2. LINKING TO TRADE STATISTICS

International trade in goods statistics (ITGS) records physical movement of goods between countries. In European Union, the international trade contains both Intra-EU and extra-EU trade.

Intra-EU trade statistics record the movement of goods between EU Member States. By convention, goods entering the Member State are called arrivals (intra-EU imports) and those leaving the territory are dispatches (intra-EU exports). Intra-EU trade is based on Intrastat data collection system. The data is collected directly from the traders. Intrastat system is closely interlinked with the VAT system relating to intra-EU trade, namely to ensure the completeness and quality of the statistical data. A system of thresholds is established to simplify data provision and reduce the overall burden on traders. For intra-EU trade, the providers of statistical information are all taxable persons reporting an intra-EU transaction whose annual trade value exceeds a certain threshold. This threshold is known as the 'exemption threshold' in the Intrastat system. Each Member State fixes its national threshold, applicable separately to arrivals and dispatches, based on a coverage rate defined by regulation of the EU.

Extra-EU trade statistics record goods imported and exported by the EU from and to non-EU countries (note that movements of goods 'in transit' through a Member State are not recorded). Data on trade in goods with non-EU countries are collected by customs authorities and are based on the records of trade transactions in customs declarations. The data collection system covering extra-EU trade is called Extrastat system. It corresponds to administrative data from the customs declarations lodged by natural or legal persons in the customs administration.

Trade of Goods by Enterprise Characteristics (TEC) is a specific dataset compiled by EU Member States on the annual basis and it describes the trade in goods between countries from the viewpoint of the enterprises. It is based on ITGS data, which has been linked to business characteristics indicators (from SBS and SBR). The trade in goods between countries is broken down by economic activity, the size-class of enterprises, trade concentration, geographical diversification and products traded.

In addition to the trade in goods, there is also international trade in services statistics (ITSS). Similarly to trade in goods, the trade in services statistics is also disseminated by enterprise characteristics (STEC). The trade in services and STEC are excluded in current microdata linking project.

### 3.2.1 LINKING TO TEC OR ITGS DATA

There are two options to link SBS population to trade data: linking to TEC data and linking directly to ITGS data. Countries may choose which option is more suitable for them considering the national circumstances and data processing practices.

The advantage of linking to TEC data may be that estimated trade amounts in Intrastat data are already allocated to individual enterprises based on their VAT data. Another advantage may be that the trade value of VAT groups is already allocated to the individual members of the group.

### 3.2.2 NON-LINKING TRADE

It is not possible to link all trade value to SBS population. Normally the trade value of ITGS/TEC should be higher than the trade value of SBS population because number of traders are missing from SBS for different reasons. This includes non-established traders, private individuals and traders with NACE activities not covered by microdata linking.

Non-established traders are foreign companies that carry out trade transactions in reporting country. They may have registered for VAT in reporting country in case of intra-EU trade. In case of extra-EU trade they may use their native VAT number or specific customs ID number

provided by customs authority. Normally the SBS population does not contain non-established traders and therefore their trade data remains non-linkable. In case the non-established trader has resident affiliate the trade could be allocated to affiliate if applicable – the resident affiliate is producing/using the goods traded by non-resident parent company.

Private individuals are also involved in international trade in goods. They must report customs declarations for extra-EU transactions and may consequently appear in ITGS data. As the SBS population covers only enterprises, the trade amount of private individuals remains non-linkable. The SBS population only covers NACE activities from B to N. The trade transactions carried out by enterprises with different NACE remains non-linkable with the SBS population.

Comparison of trade data and SBS population may indicate that there are traders who should belong to the SBS population but for some reason are missing. One possible reason could be that the trader is considered inactive according to SBS methodology but has still carried out trade transactions. As the SBS population is the basis for this microdata linking project the activities of such traders remain as non-linkable data.

An example based on Swedish trade data in 2016 illustrates the issue of non-linkable trade. About 15.7 percent, respectively 13.2 percent of all importing- respectively exporting enterprises is not captured within the SBS population, i.e. this means about 8 638 enterprises in ITGS do not match. Around 1.8 percent and 0.8 percent respectively of the no match rate above is due to the existence of so-called sub-units in the SBS survey, units that act as sub-units to the parent company but themselves not report the values for the income statement and balance sheet. The sub-units are industry-specific activities - kind of activity units and they do not follow the same statistical unit as enterprise in SBS. From the remaining part 10.7 percent, respectively 13.1 percent constitute trade from non-established traders; they are foreign companies that carry out trade but are not registered for VAT in Sweden.

If we delimit the population of traders/occasional traders to those only operating within the SBS industries B-N, then 1 658 enterprises do not match and about 11.6 percent of export and 10.7 percent of the total import cannot be attributed to the SBS enterprises.

*Note 8 – A Swedish example of non-linkable trade.*

### 3.2.3 ESTIMATED TRADE AMOUNT & USE OF VAT DATA

ITGS data may contain estimations made to replace the data losses caused by non-collected data. The share of estimations can be particularly important in intra-EU trade due to the data losses caused by non-response, partial response, reporting delays and trade below exemption threshold impacting the Intrastat data collection system. The estimations are typically allocated to product codes and partner countries in the disseminated trade statistics but not to individual enterprises.

For this reason it may be necessary to use VAT data to replace the missing Intrastat data, for the enterprises below the Intrastat threshold (or in cases of non-response). VAT data (value of goods) could be linked to individual traders and used to estimate trade below threshold as well as non-response.

### 3.2.4 VAT GROUPS AND TRADING HOUSES

A VAT group is a group of related enterprises that is treated as one entity for VAT purposes. A VAT group may report consolidated VAT reports and Intrastat declarations for all the group members. This creates challenges in linking to trade data because the data of several enterprises in SBS population may correspond to the trade data of a single VAT group.

It is recommended to allocate the trade values of a VAT group to appropriate enterprises.

Methods to allocate the declared values to appropriate statistical units can differ, depending on the reporting system used by particular countries. If, for example, the tax authorities publish information on VAT groups Intrastat data could be compared with the tax information of the VAT group. Other methods could be supplement forms for VAT group declarations that contain information concerning each VAT group member. Another approach could be to use the enterprise with the activity code that most closely matches. In cases where more than one NACE codes are applicable for a VAT number, the NACE code which is dominant for the group could be chosen.

If the VAT groups exist in ITGS microdata but their trade value is already allocated between group members in TEC microdata it may be reasonable to use TEC microdata instead of ITGS microdata in microdata linking.

ITGS data is often linkable to SBS data, but this approach can have limitations. For example, there can be a separate trading entity that carries out the trade on behalf of the producing companies. Such trading houses do not themselves have any ownership of the goods, they just organize the trade. Still they are registered as the trader in ITGS data. The trading entities typically have low number of employees, but very high trade. It is therefore often desirable to redistribute the trade to the real owner of the goods, whenever possible.

## 3.3. LINKING TO BUSINESS DEMOGRAPHY

Data on the population of active enterprises obtained from Business Demography is aligned, but not in all countries exactly identical to data obtained from SBS. This is due to the different methodologies used in defining the populations and collecting the data. In some countries, Business Demography uses different reference period than the SBS for defining the population of active enterprises. This results in some enterprises not existing within the SBS population, while existing in the BD population.

To circumvent issues arising from these differences and ensure coherent output, the SBS population was chosen to be the base population for data linking. Enterprise characteristics are obtained from the SBS if available, and otherwise from BD. These are age, survival and high-growth status.

Special attention is needed for the age variable. It is not present in the BD data as such. The associated variable which is available in BD is survival of newly born enterprises. The age-groups are then determined on basis of the year of enterprise birth. When an enterprise is in the enterprise birth group of the years t-5 to t and has survived to t, then its age is determined on the basis of the year of birth. Therefore, enterprises that are not young according to the BD, are defined as old. New enterprises, which are the result of legal restructuring, mergers, split-offs, etc., are not considered births in BD. The outlined approach automatically classifies these enterprises as older than 5 years.

The leading dataset in this Microdata linking project is the SBS dataset. This implies that in the MDL-dataset the breakdown of the population of enterprises which are – for example - 3 years old into NACE or size class groups is not necessarily the same as the population of enterprises which survived for 3-years as obtained from BD. This is because in the MDL dataset the characteristics are obtained from SBS in the current year, whereas the BD characteristics are obtained from t minus 3 and these may have changed between t and t -3.

## 3.4. LINKING TO FATS

IFATS, also called Inward FATS, is a subset of SBS and is typically based on administrative sources (i.e. the SBR which should include information on control-links between enterprises). In most cases the control-link information stems from commercial providers, which in turn are validated by SBR and IFATS producers. All in all, IFATS contains the SBS enterprises which are flagged as foreign-controlled. Linking SBS to IFATS is thus trivial.

FATS regulations require countries to compile OFATS, also called Outward FATS, with a geographical breakdown limited to extra-EU affiliates. That is, countries are not obliged (but encouraged) to compile intra-EU affiliates. Some countries include intra-EU, others do not. Also, countries should apply the ultimate controlling institutional unit approach (UCI) i.e. countries should only compile data from domestically-controlled enterprises to avoid double-counting at the EU-level. However, some countries collect data for all domestic enterprises with foreign affiliates. Further, although countries are recommended to have no cut-off thresholds, most countries apply various thresholds, e.g. only surveying enterprises with more than 20 employees. Lastly, a number of countries base OFATS on the EuroGroup Register (EGR) see note 10.

Consequently, how OFATS is produced differs significantly between countries. This leads to reduced inter-country comparability and no 'one-size-fits-all' linking methodology. The textbox describes two approaches – the Danish and a stylized EGR approach.

In Denmark, OFATS is a census survey that encompasses all (known) *legal units* with affiliates abroad and applies no thresholds. Furthermore, the UCI approach is not used in the compiling phase (but applied post-hoc in the data transmitted to Eurostat). The Danish OFATS thus contains data for all *legal units* incorporated in Denmark with affiliates abroad. Consequently, linking OFATS to the business group register and SBS in the Danish context is trivial.

*Note 9 – A Danish example on linking OFATS to the SBR.*

In principle, the EuroGroup Register contains information about all enterprises which belong to a Business Group operating within the EU. In other words, it contains the same (or more) information as the FATS statistics, and is envisioned to work as the backbone of FATS statistics by 2020. Some countries already use the EGR to produce their FATS statistics, and since the register contains the same national ID-numbers as the SBR, linking the EGR with SBS is simple.

However, issues such as low coverage (particularly for extra-EU affiliates), large year-on-year fluctuations and timeliness means the EGR by many countries is deemed to be immature and still not ready for statistical production.

*Note 10 – A stylized EGR approach to linking OFATS to the SBR.*

# 4 Validation and imputation

This chapter deals with several aspects of the output phase of the project, including internal and external data validation as well as issues related to imputation.

## 4.1. VALIDATION

The purpose of this BD-TEC microdata linking project was to provide breakdowns of the SBS population by coupling data from other sources namely Business Demography, TEC and FATS. Only in an ideal world there would be no differences in the populations of these sources and of the characteristics of all elements in these populations. Thus, in practice, the earlier published results are not expected to be identical to the totals in the dataset of linked populations.

As the published variables are all SBS variables, (and data from other sources are only used for breakdowns) the SBS population was the central population (including the enterprises' characteristics therein). This ensures that published SBS data are, as closely as possible, consistent with microdata linked output.

Data validation of the SBS data consists of the following elements:

- **Internal validation**

  Internal validation of the data sets checks whether the microdata linked tables are consistent between each other. It involves two elements:

  - **Inter-table validation**

    Totals of each category is checked to be identical to the sum of the elements contributing to this total

  - **Intra-table validation**

    When tables have at least one explanatory or spanning variable in common and have the same response variable the outcome in the set of tables for these cells should be identical.

- **External validation**

  The earlier published totals for the SBS variables are checked to be identical to the totals. This validation is only performed where the design of the tables allow this comparison. In the current project the comparison can only be made between tables 1 and 3, due to double counting (by design of the table) in table 2.

  Comparison with totals from FATS, TEC, BD is performed, with the knowledge that differences to some extent can be explained as outlined in the previous chapters.

*Note 11 – Validation strategy for SBS data.*

Data validation of trade data is another domain. A possible starting point is to check the ratio between exports and turnover in matching enterprises. Furthermore investigating missing matches both for imports and exports to the SBS population on enterprise level is a good practice.

Other approaches include:

- With large export ratio or wrong industry; the solution is to reassign values between enterprises within the enterprise group
- For agriculture exports or other export outside the SBS population there are enterprises with few employees handling exports. For export figures for those enterprises belong to an enterprise group and trade has been redirected to other enterprises within the group that belong to SBS.
- For some of the enterprises with high imports or exports figures that are not matched to SBS and belonging to enterprise groups, an effort should be made in order to find another enterprise within the enterprise group that is included in SBS, to allocate the ITGS data.

*Note 12 – Validation strategy for trade data.*

Below are Norwegian country practices for some common problems when combining ITGS data with SBS data.

- **Redistribution of International Trade in Goods Data using enterprise group information**

  In the validation process of Norwegian data for the BD TEC Microdata linking project Statistics Norway has to some degree used information about the enterprise group to redistribute Trade in goods statistics (ITGS). This has been done where Statistics Norway has been unable to match the enterprise used in trade data with a corresponding enterprise in SBS; but where another enterprise was found within the enterprise group that is included in SBS. In trade data, enterprises used are often not the same as the "producing" enterprise from SBS. The redistribution has been done by checking data and then creating an algorithm that redistributes the described data.

- **Redistribution of fish (or other agricultural products)**

  Fish is a product that is extensively sold abroad by independent trading companies, often with two- digit NACE code 46. The enterprises containing the turnover and employees behind these products have either industry NACE 03 or 10. Due to this issue discrepancies appear between ITGS- and SBS data. The turnover and employees are typically placed in different enterprise in SBS and ITGS data. Since fish constitutes a large part of the Norwegian economy it was important to correct these data. Without this redistribution Statistics Norway would get very odd figures for enterprises in these categories.

- **Enterprises not part of an enterprise group**

  Many enterprises in ITGS are lacking from SBS and in addition are not part of an enterprise group. These enterprises cannot be linked using the enterprise group method described above. Statistics Norway has used another approach which involved searcing among the owning enterprises. In some cases they found a company belonging in the SBS in the ownership structure of the trade enterprise lacking from the SBS-survey, and have then redistributed the trade to the enterprises that are part of SBS.

- **Manual redistribution of enterprises not part of an enterprise group**

  In the most severe cases, where large enterprises have conflicting information regarding turnover, employees and trade – Statistics Norway have manually redistributed trade from the trading entity to a more appropriate enterprise. This has only been done where they are able to establish a kind of ownership/connection between the enterprises.

*Note 13 – Norwegian approach to issues when matching ITGS with SBS.*

## 4.2. MISSING DATA AND ALTERNATIVE SOURCES

Missing data in the context of microdata linking can be handled in various ways. The first way to deal with missing data is to see if there is an alternative source for the missing data. The following table describes the various sources for selected domains where missing data could be present.

- **Structural Business Statistics**

  Annual accounts for turnover and number of employees.
  Taxation data for turnover, purchases and number of employees.

- **Business Demography**

  Taxation data for knowing which and when units paid taxes or VAT.
  Customs data for knowing which and when units exported or imported.
  Legal Business Register for information on liquidations and bankruptcy.
  Statistical Surveys such as short term statistical surveys or the special survey for all newly registered units.

- **ITGS/TEC**

  VAT system for intra-EU export and import values.
  VIES for intra-EU export value only.
  Annual accounts.

*Note 14 – Alternative sources for various statistics.*

Country practise: In STS domain, Statistics Finland controls for mergers and split-offs in a set of enterprises to obtain organic growth rates. The data that indicates restructuring event is sourced at Tax office data, which is delivered monthly to Statistics Finland. Events can also be spotted during manual editing of largest enterprises by statisticians.

Assume that firm 1 is examined after an event (merger or split-off) where N firms are involved. Then the estimated employment of firm 1 one year ago is calculated by

$$EMP_{firm\ 1,t-12} = EMP_{firm\ 1,t} * \frac{EMP_{firm\ 1,t-12,firm\ 2,t-12,\ldots firmN,t-12}}{EMP_{firm\ 1,t,firm\ 2,t,\ldots firmN,t}}$$

where t is the time periods in which the adjustment is computed, and N is the number of firms involved in a merger or split-off. The sum of the previous year employment levels in all the firms involved in the event is divided for each continuing firm weighted by their relative size at present time t. Let us go through some simple numerical examples to see how this works:

1) Assume a firm A with 2 employees in period t, that had 1 employee in t-12. Firm A acquires firm B with 1 employee at time t, and had 1 employee one year ago. Firm A, which continues existing, will be assigned a new estimated number of employees for the comparison year, in order to make the growth rates comparable year-on-year. The comparison values of firm A is estimated as2(1+1)(2+1)= 4/3, and the rate of change for A becomes (2 + 1)/(4/3) = 2.25 (as opposed to 3 if no correction is done)

2) Consider the situation where firm A is split into smaller units, say B and C. A has 3 employees at time t−12, B has 3 employees at t and C has 2 employees at t. B and C did not exist at t−12, so their comparison values become:(3/3)*3 = 3 and (2/3)* 3 = 2, resulting in the rate of change for B and C to be 3/3 and 2/2 (equal to1 for both firms). In essence, the growth rate is forced to be the same among the continuing firms after a split-off.

In the context of microdata linking, the data from this procedure is used by Statistics Finland to validate BD data. Of course this data could be used in later work to produce indicators in a more timely fashion. As a side note, discrepancies in data quality even inside an NSI between statistical domains is also an important lesson of microdata linking exercises. Most often the stove pipe approach of official statistics means that the statistical production systems are separately developed for each statistics, so that there may be different ways of dealing with the same problem even inside the same office.

*Note 15 – Country practice: In Finland, Short-term business statistics growth correction methodology can be used to identify cases where resources are transferred between enterprises.*

## 4.3. IMPUTATION AND MISSING DATA

If there is no alternative source for the missing data, imputation is another option. The following sections describe some approaches to imputation used in this BD TEC microdata linking project.

In the BD TEC this microdata linking project the typical target quantities to be estimated can be expressed as table of totals or a table of counts for a reference (sub)population. Missing data can usually not be ignored because they are not random. When overlaps between datasets are used to analyze relations between variables this may result in small number of observations and thus large uncertainties.

Two methods to deal with missing data are discussed below. The choice for a certain method may depend on missingness patterns, or availability of auxiliary variables.

The methods described here are Repeated weighting and Mass-imputation. There are several advantages and disadvantages related to the choice of imputation model:

- **Repeated weighting**

  + Microdata consistencies, meaning the relations between variables of the same unit are maintained.
  - Not suited for estimation of small subpopulations.
- **Mass-imputation**

  + Makes good use of all information in the linked data.
  + Flexible and can be adapted for different variables.
  + Suitable for small subpopulations.
  - Cumbersome to specify and check imputation models.
  - Biased when analyzing relations which are not sufficiently accounted for in the imputation model.
  - Relations can become spurious because large numbers of missing data make it impossible to estimate the - model parameters accurately.

*Note 16 – Advantages and limitation of repeated weighting and mass-imputation.*

### 4.3.1. REPEATED WEIGHTING

Repeated weighting is a technique that was developed at Statistics Netherlands to solve the problem of inconsistencies among tables of estimates based on multiple data sources.

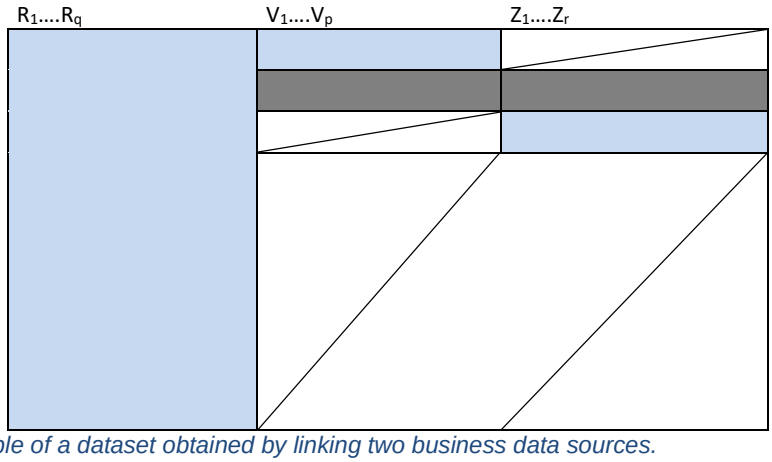



*Figure 1 - Example of a dataset obtained by linking two business data sources.*

Figure 1 shows the situation of two datasets linked to the backbone register, the SBR in our case. Grey parts are observed and white crossed blocks are missing. Variables R are available in the BR, variables V in the first linked dataset and variables Z in the second dataset. The dark grey colour indicates the units, for which both V and Z are observed, i.e. the overlap between the two datasets. This is a realistic scenario for business datasets as most such datasets do overlap to some extent, because large enterprises are typically completely observed in take-all strata. But this also illustrates that such overlaps can be highly selective, especially with respect to size-class.

If we are interested in the relation between V and Z variables, let's say a cross-tabulation V x Z of categorical V and Z variables, the Repeated weighting strategy proceeds as follows. For convenience we assume that the target population is the full set of units in the SBR. Start with estimating the V and Z marginal from their respective sources. These may be obtained by a standard weighting procedure, where weighting variables are selected from the R variables. This is the usual way official figures on population totals or means are obtained from a single data source. Next, compute initial estimates for the V x Z table based on the dark grey block. This can be accomplished by a standard weighting procedure as well, although now one should be even more careful to find a suitable weighting model that reduces selectivity bias. The V and Z marginal of such a table will not agree with those already computed, since the latter are based on more information (more observations). In order to correct for this, the table of initial estimates is recalibrated so as to agree with the previously computed V and Z marginals. This can be achieved by reweighting the darker grey block with respect to V and Z variables, using the previously computed V and Z marginal as control totals. This results in adjusted estimates for the V x Z table whose marginal agree with those based on the separate data sources. This procedure generalizes to more than two data sources.

### 4.3.2. MASS-IMPUTATION

Imputation is different from weighting. Instead of assigning weights to units for which variables are observed, the focus is on the missing units. For the unobserved units the values of the variables are filled in. Imputation is often used to fill in item non-response in surveys. If imputation is used to impute values that are missing by design then the term mass-imputation is used.

The imputations are generated by means of an (explicit or implicit) imputation model. Mass imputation makes more use of auxiliary variables for developing imputation models compared to survey weighting (including repeated weighting) and can be made robust to outliers and other potential errors in the covariates. After mass imputation, estimates for population totals can be obtained by simply counting or summing the values of the corresponding variables. In this case the population is the entire SBR population.

Mass imputation is often based on advanced imputation methods that satisfy edit and calibration restrictions on the data, for a more extensive explanation, see footnote1. Such an approach may be useful if it is important to maintain numerical consistency with previously published official population totals.

The mass imputation approach relies on the ability to capture all relevant variables and relevant relations between them in the imputation model, to estimate the model parameters sufficiently accurately. Given that all relevant variables and relations among them can be captured accurately by the imputation model, the approach is very straightforward. When the relationships between variables also hold for sub-groups of the population it is also

[1] Pannekoek, Shlomo, & Waal, d. (2013). *Calibrated imputation of numerical data under linear edit restrictions.* Retrieved from https://projecteuclid.org/euclid.aoas/1387823307
Kim, Reiter, Wang, Cox, & Karr. (2014). *Multiple Imputation of Missing or Faulty Values Under Linear Constraints.* Retrieved from https://amstat.tandfonline.com/doi/abs/10.1080/07350015.2014.885435#.W1CCFbh9iUk
Waal, d., Coutinho, & Shlomo. (2017). *Calibrated Hot Deck Imputation for Numerical Data Under Edit Restrictions.* Retrieved from https://academic.oup.com/jssam/article/5/3/372/3797270

straightforward to compute totals for these sub-groups summing up the values for the specific variables of these sub-groups.

A fictitious example of the last item is the following. Suppose we couple a survey on 'Spending of Households' and a survey on 'Household situation'. The first contains information on spending on dogfood, the second on possession of a dog. When these data-sets were combined and mass-imputated the variable 'possession of a dog' was not considered relevant enough for a breakdown of the household spendings, and was not incorporated in the imputation model. Therefore, a non-zero value is imputed for 'spendings on dog-food' for households without a dog.

When a researcher, without detailed knowledge of the imputation method, uses this dataset he observes that households without a dog spent substantial amounts of money on dog-food. He concludes that there is apparently that much poverty that people have to eat dog-food.

*Note 17 – Fictitious example of the risk of mass-imputation methods.*

### 4.3.3. COUNTRY PRACTICES

The following section contains examples of some country practices when it comes to imputing the SBS survey further. It also shows how to effectively use administrative data to avoid surveys and imputation.

In the Netherlands, the population of the SBS survey consists of approximately 50 to 60 thousand enterprises, with a total population of approximately 1 million enterprises. The goal is to determine values for the non-survey units for the main variables such that the totals are consistent with published totals.

The output of the SBS production process in the Netherlands consists of a dataset in which for each enterprise in the survey population a value is assigned to the SBS variables. A survey weight is determined at the enterprise level.

Then, the SBS published total for variable $V_{i,total} = \sum_{j=1}^{n} survey\ weight_j \times V_{i,j}$ (1)

where *n* is the number of enterprises in the survey

In order to calculate a value for each variable for each enterprise in the total population, this total population is divided in -supposedly homogeneous - groups of NACE 3-digit *x* size-class groups.

Using formula (1) the total $V_{i,total,\ k}$ for group *k* can be calculated for each variable $V_i$.

Subsequently the gap between the sum of the measured values (without the survey weights) and the SBS total is distributed over the non-survey population based on the relative share of persons employed. In a formula: the value of variable $V_i$ in group $k$ of non-survey enterprise $E_l$ equals:

$$V_{i,k,l} = \frac{persons\ employed_l}{\sum persons\ employed_{non-survey\ units,k}} \times \left( V_{i,total,k} - \sum_{j=1}^{n} V_{i,j,k} \right) \quad (2)$$

The number of persons employed is used as auxiliary variable and is available at the micro-level from the Social Security database which contains data on all employees. The number of owners per enterprise added to obtain the total number of persons employed.

Consistency between output variables is maintained as each variable is multiplied by the same factor. A disadvantage is that all non-survey enterprises in the same group have the same performance, for instance turnover per employee. Characteristics used in this study such as international trade or foreign ownership that describe international orientation of an enterprise do thus not influence performance of non-survey enterprises which are in the same group.

Current research focuses on the use of different auxiliary variables based on tax-information for different groups of SBS variables, while maintaining relationships between variables.

*Note 18 – Country practice: Mass-imputation in the Netherlands.*

In Denmark the SBS population is around 210.000 enterprises. The sample for the surveys covers 4 % of the SBS population and about 72 % of the total revenue of the SBS population. Besides the sample the survey is enriched by administrative data from the Tax and Customs Authority, this data covers an additional 43 % of the SBS population and around 16 % of the total SBS revenue. The remaining enterprises in the population, around 53 % of the enterprises and 12 % of the total revenue are imputed at enterprise level.

The imputation is implemented in the following way:

1. Stratification averages are calculated for each strata (approx. 525 strata), firstly for the surveyed enterprises. The averages are trimmed to avoid influence of extreme observations.
2. The survey stratum averages are used to enrich and impute the data from the Tax and Customs Authority for variables and or units that are not reported to the Tax and Customs Authority.
3. After the enrichment and imputation in step 1 and 2 the stratum averages are updated.
4. The final stratus averages are used for imputing the remaining population.

Currently methodologies are being developed which utilize XML data reported by enterprises to the Danish Business Authority. This data provides a third source of information which will make more precise imputation possible.

*Note 19 – Country practice: Imputation using administrative data and strata averages in Denmark*

In Portugal the integrated business account system (IBAS) is the basis for the production of structural business statistics.

- IBAS integrates 3 main sources of information:
- The Simplified Business Information (SBI), which must be presented by all companies
- Data for sole proprietors and independent workers received from Tax Authorities
- Business Register (BR) data.

Comparisons with other sources such as Lists of Personnel and specific surveys are made in order to calculate the population of active enterprises that should be considered each year.

Theoretically, SBI covers all population of companies; however, not all companies deliver their statements on time to be incorporated in the production of SBS statistics. Missing statements for a given year doesn't mean that the company didn't have activity in that year. In this context, there is the need to determine, for each year, the number of companies that have to be added to IBAS population. These companies are considered in the scope of IBAS with data based on BR.

In the IBAS data base there is information for more than 1 million units, covering NACE sections A to S, except O. After the identification of the active enterprises, results are obtained through a wide range of variables, around 500. These are built through the appropriation of administrative data (observed variables) for almost all the population of companies. However, every year, a set of companies have to be estimated because they do not submit their data timely for their integration in IBAS.

*Note 20 – Country practice: SBS statistics using administrative data sources in Portugal.*

## 4.4. GENERAL CONSIDERATIONS

The results from microdata linking can differ from the officially published totals for many reasons.

Nonetheless, the first step is to check what the published totals are in the SBS releases. The deviations of MDL output from this total should ideally be well understood.

An obvious deviation can occur if the data has accumulated in the registers after the official release has been made. It is a good practice to put a disclaimer that explains the reason for the revision if one uses the updated data. There may be good reasons not to use the updated versions of the data, such as revision policy or technical issues.

There might have been notable errors in the original data sources that are revealed during the linking process. This can be due to more detailed breakdowns, or more careful inspection of the data. In this case, it is advisable to make the necessary corrections to the data up to the level that any remaining errors do not hamper the economic message that the publication conveys. An important case on this point is the analysis of young enterprises using Business demography statistics and SBS statistics. If a very large enterprise has been wrongly classified as young due to a legal restructuring, it will seriously hamper the economic message that is given by such breakdowns. A good practice is to rank young enterprises by size and verify that the largest ones are correctly classified. An important additional variable that should be used in this case, is the enterprise group link and the identity of the mother.

It is a good idea to select a few well known enterprises, after having carried out the linking exercise, to make sure that the resulting data correctly describes the activities of the known enterprises. In small open economies, there are often a small number of very important enterprises that are good candidates for this check.

More generally, it is recommended to:

- Be mindful of the probability of having an error in certain breakdowns, and also think ahead how this might affect the analysis in order to decide how much should be invested in the inspection of the data.
- Apply economic reasoning to verify that the linking is appropriate for the analytical goal. For example, the data which is based on a sample of large enterprises (for example the International trade in services) is not appropriate for an analysis which tries to compare the different behavior of large and small enterprises.
- Before disseminating, try to discuss with subject matter experts to confirm that the results make economic sense. These are new breakdowns, and it may not be easy to spot mistakes on your own.

# 5 Dissemination strategy, confidentiality and revisions

## 5.1. CREATING BREAKDOWNS

Creating breakdowns is a cornerstone in the process of creating tabular output for analysis and dissemination. The breakdowns or categorizations used in the project, which have been proven to be useful in other areas of business statistics, are the following:

### 5.1.1. ECONOMIC ACTIVITY

The data in this project is compiled over several years; this means that careful consideration should put into classifying the Economic Activity of an enterprise, the main consideration should be whether it should be the activity in the birth year of the enterprise or the reference year of the compilation. The experience from this project is that the reference year provides the most comparable longitudinal data.

**Economic Activity based on NACE Rev 2. Groups**

- Manufacturing – Nace section C
- Wholesale trade – Nace section G45 & G46
- Other – Nace section B-N, S95, excluding C, G45 & G26 and K

*Note 21 – Activity breakdowns.*

### 5.1.2. SIZE CLASSES

Defining size classes for the enterprise is also a complex task. For statistical purposes, SMEs are generally defined as those enterprises employing fewer than 250 persons. The number of size-classes available varies according to the activity under consideration. However, the main groups that are often used for analytical purposes and presenting the data include:

The European Commission Recommendation (2003/361/EC), adopted on 6 May 2003, classifies SMEs with two additional constraints: annual turnover larger than 50m €, balance sheet total larger than 43m €.

An enterprise which is part of an enterprise group may need to include data on persons employed, turnover and the balance sheet from the whole group. There are specific guidelines on how much of the group's employment/turnover/balance sheet should be included to determine the SME status.

These guidelines are complex and are therefore difficult to use in statistical systems. This is the reason, why for statistical purposes size class information is only based on the number of

persons employed in the enterprise itself, without looking at the turnover, balance sheet or employment data from the group that the enterprise belongs to.

However, doing this has consequences for analysis and policy, since enterprises belonging to a domestic enterprise group may be different from independent enterprises, for example in their ability to access finance, their bargaining power, possibilities to expand to foreign markets, and various other aspects of doing business.

The number of standard size-classes available within SBS varies according to the activity under consideration. However, the main groups that are often used for analytical purposes and presenting the data include:

**Size class based on number of employees on legal unit level**

- SME – Between 0 and 249 employees
    - Micro – Between 0 and 9 employees
    - Small and Medium – Between 10 and 249 employees
- Large – More than 250 employees

*Note 22 – Size class breakdowns, based on persons employed.*

### 5.1.3. TYPE OF OWNERSHIP

Microdata linking output contains a breakdown of the population based on a combination of size class and type of ownership (controlling links). This information can be derived from the SBR, and often also by the EGR and FATS information.

The MDL project has introduced various categories based on enterprise dependency status inside an enterprise group. The controlling links are typically found in SBR between the domestic mothers, the entity that has 50% or more of the voting power in the enterprise, and the affiliated firms. Microdata collected for Foreign affiliates statistics (FATS) include information on whether an enterprise is controlled from abroad (IFATS), or an enterprise has affiliates abroad (OFATS). The EuroGroup Register (EGR) contains similar information and can be used as an alternative to FATS and can additionally provide information on the total size of the controlling institution.

For full control, the ultimate controlling institution needs to hold at least 50% of the votes, which is the type of control we have restricted our attention to in the microdata linked tables.

Based on controlling relationships, an enterprise can be allocated to the following categories:

**Group status based on domestic relations (BR), outward relations (OFATS) or inward relations (IFATS)**

- Independent – Enterprise neither controlled nor being controlled by another enterprise
- Dependent – Enterprise is controlled by or controls another enterprise
    - Dependent – belonging to a domestic group
    - Dependent – belonging to an international group
        - Dependent – Being controlled by a foreign enterprise
        - Dependent – Having one or more foreign affiliates

Summing the number of employees on the enterprise group level utilizing the breakdowns described above makes for another interesting breakdown.

*Note 23 – Group status definitions.*

### 5.1.4. TRADER TYPOLOGY AND PARTNER COUNTRY

Characterizing trader typology is done using two breakdowns, one focusing on export only and another which focusses more broadly on trader behavior.

**Simple trader typology, based on TEC (or ITGS)**

- Exporter – Enterprises that have more than 5000 euro in exports and that have export intensity (ratio between exports and turnover) larger than 5 %.
- Non-Exporters – Enterprises below the threshold

**Complex trader typology, based on TEC (or ITGS)**

- Traders
    - Exporter only – Enterprise has more than 5000 euro in exports and export intensity larger than 5 % but are not importers
    - Importer only – Enterprise has more than 5000 euro in imports and import intensity larger than 5 % but are not exporters.
    - Two-way-traders – Enterprise satisfies threshold to be both exporter and importer.
- Occasional trader – Enterprise has export or import but below the value or intensity threshold.
- Non-trader – Enterprise has no export or import

*Note 24 – Trader typologies.*

It is worth noting that there can be large amounts of export and / or import values among enterprises below the thresholds. As an extra validation, it is best practice to examine the largest cases, since ratio between turnover and export can be deceiving. Besides the trader typology, partner country is also broken down:

Partner country, based on ITGS (or TEC)

- EU28 – All 28 EU countries
- Extra EU
  - USA and Canada
  - China (including 'CN', 'HK', 'MO' & 'TW')
  - Rest of world

*Note 25 – Partner countries.*

### 5.1.5. DEMOGRAPHY AND GROWTH

In this project growth rates of enterprises were also considered important. To capture fast growing enterprises the following breakdown was created.

**High growth definition**

- High growth enterprises – Enterprise has 10 or more employees in the start of a three year period and has an average annualized growth in employment greater than 10 % over the three year period.
- Non-High growth enterprises – Enterprises below the threshold

*Note 26 – Definition of high-growth.*

The last breakdown in this project was creating to capture the demographic characteristics of the enterprise:

**Age, based on Business demography**

| Breakdown | Age | t-0 | t-1 | t-2 | t-3 | t-4 | t-5 |
|---|---|---|---|---|---|---|---|
| 0-1 years | 0 years | 1 | | | | | |
| | 1 year | 2 | 1 | | | | |
| 2-3 years | 2 year | 2 | 2 | 1 | | | |
| | 3 year | 2 | 2 | 2 | 1 | | |
| 4-5 years | 4 year | 2 | 2 | 2 | 2 | 1 | |
| | 5 year | 2 | 2 | 2 | 2 | 2 | 1 |
| Other | | | | | | | |

The numbers in the tables are the so called BD codes: Where 1 - the enterprise is born in the reference year, and 2 – the enterprise is active in reference year. All enterprises 5 years or younger are considered young enterprises.

*Note 27 – Definition of age.*

## 5.2. SELECTING VARIABLES

The following variables provide a solid and comprehensive foundation for analysis and dissemination. Almost all variables, besides the trade values variables are from the Structural Business Statistics.

**Selected variables, based on SBS and ITGS**

- Number of enterprises
- Number of employees
- Number of persons employed
- Export value (in euros)
- Import value (in euros)
- Turnover
- Purchases of goods and services
- Employment in FTE
- Wages and salaries
- Value added at factor cost

*Note 28 – Selecting variables.*

## 5.3. SUGGESTED OUTPUT TABLES

This chapter contains the BDTEC tables to be requested by Eurostat from Member States in excel format[2] and the possible indicators to be created based on those tables. The data requested in the tables is directly obtained from the data series defined for this project.
The notes below summarizes the output breakdowns and variables used in the recent BDTEC exercise. The rationale for choosing these breakdowns are further elaborated in previous chapter 5. Particular focus in the chosen breakdowns was to provide more information on age structures of enterprises, including involvement in trade.

| TABLES | BREAKDOWNS | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Nace | TypeTrader1 | TypeTrader2 | Age | Growth | SizeClass1 | PartnerLocation | GroupStatus |
| Table 1A | | YES | | YES (levels 1,2,4) | | | | |
| Table 1B | | | YES | YES (levels 1,2) | | | | |
| Table 2A | | YES | | | YES | | | |
| Table 2B | | YES | | | | YES (levels 1,2) | | |
| Table 2C | YES | YES | | YES (levels 1,2) | | | | |
| Table 2D | | YES | | YES (levels 1,2) | | | YES (levels 1,2) | |
| Table 3B | | YES | | | | YES (levels 1,2) | | YES (levels 1,2) |
| Table 3C | | YES | | YES (levels 1,2) | | | | YES (levels 1,2) |

*Note 29 – Summary of breakdowns*

[2] Please note that the excel sheets are attached to this guideline.

| TABLES | VARIABLES | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Number of enterprises | Number of employees | Number of persons employeed | Employment in FTE | Turnover | Value added at factor cost | Purchases of goods and services | Exports value | Imports value | Wages and salaries |
| Table 1A | YES | YES | YES | | YES | | | YES | | |
| Table 1B | YES | YES | YES | | YES | | YES | YES | YES | |
| Table 2A | YES | YES | YES | | YES | | | YES | | |
| Table 2B | YES | YES | YES | | YES | | | YES | | |
| Table 2C | YES | YES | YES | | YES | | | YES | | |
| Table 2D | YES | YES | YES | | YES | | | YES | | |
| Table 3B | YES | YES | YES | YES | YES | YES | | YES | | YES |
| Table 3C | YES | YES | YES | YES | YES | YES | | YES | | YES |

*Note 30 – Summary of variables*

## 5.4. CONFIDENTIALITY AND DISCLOSURE CONTROL

External confidentiality with SBS results is checked. Since the NACE aggregation level in this project is at a high level, it was not expected that confidentiality which originates from published SBS results have to be imposed on the results of this MDL project. This check was part of the confidentiality analysis.

Several methods for determination of the confidentiality pattern were applied. The most detailed method would be to determine a single pattern for a group of related variables. Thus for SBS variables associated with turnover – Purchase of goods and services and Value added at factor cost, the confidentiality pattern based upon Turnover is applied. For variables Number of employees, Employment in FTE, Wages and Salaries the confidentiality pattern obtained from Persons employed is applied.

No confidentiality was applied to the variable Number of enterprises. The confidentiality patterns for the variables Export value and Import value were, where applicable, independently determined.

When tables have at least one explanatory or spanning variable in common and have the same response variable, the confidentiality pattern for these common cells in the set of tables should be identical. This restriction was input for the Statistical Disclosure software.

## 5.5. REVISIONS

Revision policy establishes general principles and procedures for revisions of BD-TEC data to ensure the quality and reliability of disseminated statistics.

Regular data revisions of the previous reference years are not required. Once the output tables are compiled and delivered to Eurostat, the data is considered as final and further revisions of the same reference year are not needed under normal circumstances.

Microdata linking takes place considerable time after the end of reference year and this means that source data has been already updated and revised before it is used for microdata linking purposes. TEC data should be transmitted to the Eurostat 18 months after the end of reference year and no further revisions are required. This means that the TEC data used for microdata linking should be considered as final. However, if a significant revision of TEC data takes place after the compilation of BD-TEC datasets for the same reference year, it is recommended to revise BD-TEC data as well to maintain comparability with TEC.

Revision of BD-TEC data is required if errors have been made during the compilation of output tables. Errors may occur during the data linking process, applying confidentiality to data or production of output tables. If such error is detected, it is required to transmit revised data to Eurostat together with explanation that describes the type of error and reference periods, indicators and tables affected. The revised data should be delivered in the same format as original data.

# 6 Closing thoughts and future work

Microdata linking, i.e. linking of different statistical databases is a cost efficient way to improve and diversify existing statistics.

Issues that need to be addressed are related to confidentiality practices in each NSI, handling the enterprise dynamics/events over time, solving the issues of no-matching population and validation/consistency checks.

In the BD TEC microdata linking project, most NSIs mentioned difficulties regarding firm dynamics. Business registers in some countries do not have information on ownership changes. This means, for example, that analysis on the high growth enterprises in general is biased and should therefore be interpreted with caution.

Regarding validation checks, there has been no simple way of handling the consistency problems in the microdata. Most countries involved in this project mainly focused on the largest 100 enterprises, by checking the coherency between turnover and exports, and analysing the employment changes over time to detect any erratic demographic events. All these inconsistencies were revised before the project started prior to the analysis.

Microdata linking should be driven by well-defined policy needs (national and international). This means that the projects would be output driven and the output should be published quickly after the project. It would be better to have a focused output rather than engage in collecting an overwhelming amounts of data. Visible results would help in motivating current and future project partners and contribute to national and international policy shaping. Focusing on output would also help in the validation process of the data.

In the future, combining the SBS data to social statistics by including variables available at individual level such as educational attainment, gender, income/salaries, years of work experience and more is expected to enrich and deepen linked microdata analysis. This would allow new kind of analysis, linking business dynamics to social aspects, such as gender or income inequality.

Microdata linking provides unique opportunities to create more insight into the business economy without increasing the burden on respondents. This means that microdata linking will become increasingly relevant in the future. Looking into the future there are two main challenges than should be addressed, as described in the Note 31 below.

**Impact of FRIBS on ITGS and microdata linking**

In the foreign trade statistics there will be significant changes in the Intrastat system in the coming years. The new legislation (FRIBS) establishes a system for microdata exchange (MDE) between Member States. This means that all countries only have to collect their exports (dispatches) data and can use the partner data for the calculation of imports (arrivals) trade flow (Modernisation of Intrastat). The business ID of trading partner will be exchanged along with other micro data about trade transactions. At the moment some specifics of the new system are still unclear, as well as the precise implementation time-schedule.
The implementation of MDE could have some impact on the microdata linking. It should be possible to identify the importers according to the ID's received from partner countries, but some impacts on quality and coverage could be expected. In case of missing on non-linkable data it should be still possible to use VAT data to replace any data losses.

**Microdata linking of trade of goods and trade of services**

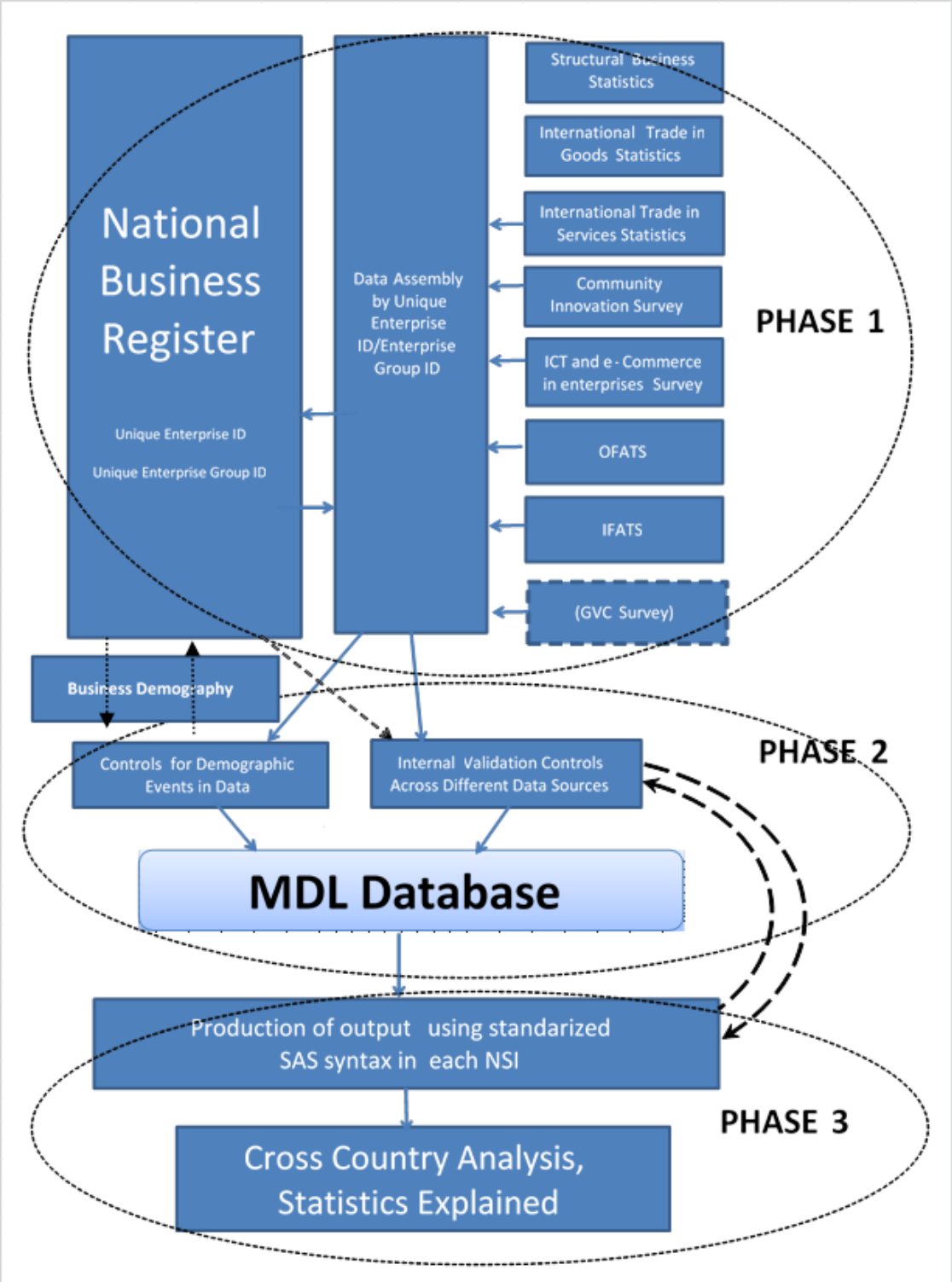
Current MDL project covers only the trade of goods and excludes trade of services. The similar projects in foreign trade statistics (TEC) and (STEC) also concentrate on their respective fields. But in practice the trade of goods and trade of services are closely connected. The same companies are often trading with both goods and services; the goods are often exported or imported with accompanying services (for example assembly or maintenance). Even the physical distinction of goods and services are sometimes difficult to determine as goods previously traded in physical format (software and music on CD's and DVD's) are now traded via internet. To get full overview and understanding of foreign trade by enterprise characteristics it would be necessary to observe trade of goods and services together.
It should be useful and interesting in the future to include ITSS data in microdata linking exercises to explore the feasibility to develop new statistical indicators and better address the needs of users.

*Note 31 – Challenges for the future.*

| Publication | Purpose | Statistical source | Findings[3] |
|---|---|---|---|
| Microdata linking – international Sourcing | Test the centrally developed methodology for linking datasets and producing standardised outputs across a number of participating countries. Produce new information on enterprises that responded to the 2009–11 international sourcing survey, carried out in 15 countries including all those taking part in the microdata linking project. | • Structural Business Statistic (SBS)<br>• International trade in goods statistics (ITGS)<br>• Foreign Affiliates Statistic (FATS)<br>• International Sourcing Survey (ISS). | Foreign-controlled enterprises are more active in international sourcing and trade.<br><br>There seems to be a negative impact on employment due to International Sourcing between 2008 and 2011. |
| Foreign affiliates statistics - employment by business function | Investigation of the employment record of foreign affiliates, by business function, of enterprises in European Union (EU) Member States. | • Foreign Affiliates Statistic (FATS)<br>• Structural Business Statistic (SBS)<br>• International Sourcing Survey (ISS) | Employment in foreign affiliates of European enterprises is falling less than in domestic enterprises.<br><br>There is no evidence of substantial movement of knowledge-intensive business functions to destinations outside Europe. |
| Statistics on small and medium-sized enterprises - Dependent and independent SMEs and large enterprises | Getting a better understanding of SMEs and their impact on employment, trade and economic activity. | • Structural Business Statics (SBS)<br>• International trade in goods statistics (ITGS)<br>• Business Register (BR) | SMEs are a very important part of the economy, as they represent around 99 % of all enterprises and employ an increasing number of persons.<br><br>Most enterprises are independent and do not belong to an enterprise group, but within the SMEs medium-sized enterprises are very often part of a group. This is most common in manufacturing and to a lesser degree in knowledge-intensive business services.<br><br>In most countries, dependent SMEs are more open to international trade than independent SMEs. |

[3] Please note that the findings are in no way inert to the changing economic and structural landscape. The findings should therefore be seen in the context in which they are produced, written and published.

| | | | |
|---|---|---|---|
| Statistics comparing enterprises which trade internationally with those who do not | The main purpose was to examine the behavior of enterprises that engage in international trade. | • Structural Business Statics (SBS)<br>• International trade in goods statistics (ITGS)<br>• Business Register (BR) | More than one in five (23 %) of all manufacturing enterprises in the eight participating countries are international traders. Denmark and Austria have the highest shares (around 40 %); Germany is a little above average (25 %) while other countries were below the average.<br><br>Foreign-controlled wholesale and retail importers account for the highest shares of imports in most countries. In 2012, in Sweden, they accounted for 64 % of total imports of goods. |
| Statistics on enterprise survival and growth prospects between 2008 and 2012 | The main purpose is to facilitate analysis of growth trends for micro, small, medium and large enterprises. | • Structural Business Statics (SBS)<br>• International trade in goods statistics (ITGS)<br>• Business demography (BD) | Survival rates are inversely related to enterprise size. On average — and across all countries — SMEs are characterised by a higher number of deaths, both in absolute and relative terms. In the SME category, micro and small enterprises have the lowest chances of survival<br><br>Continuing exporters (enterprises that exported in 2008 and 2012) tend to be the largest in size, and they are heavily represented among medium-sized and large enterprises. Non-exporters are generally less likely to change size class than enterprises engaging in exports. |


National Business Register
Unique Enterprise ID
Unique Enterprise Group ID
Data Assembly by Unique Enterprise ID/Enterprise Group ID
Structural Business Statistics
International Trade in Goods Statistics
International Trade in Services Statistics
Community Innovation Survey
ICT and e-Commerce in enterprises Survey
OFATS
IFATS
(GVC Survey)
PHASE 1
Business Demography
Controls for Demographic Events in Data
Internal Validation Controls Across Different Data Sources
PHASE 2
MDL Database
Production of output using standarized SAS syntax in each NSI
Cross Country Analysis, Statistics Explained
PHASE 3

## 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/contact

**On the phone or by e-mail**
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 electronic mail via: https://europa.eu/contact

## 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: http://europa.eu

**EU Publications**
You can download or order free and priced EU publications from EU Bookshop at: http://bookshop.europa.eu. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/contact)

**EU law and related documents**
For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex at: https://eur-lex.europa.eu

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

KS-GQ-19-013-EN-N

# Micro data linking

Microdata linking (MDL) provides an opportunity to develop new statistics and indicators and to discover new information both when using existing data sets but also when combining with new data collections.

The guidelines for MDL gather past experiences and best practices of the microdata linking. The key concepts and sources are described in broad terms. The central role of the Business register is emphasized when carrying out microdata linking exercises. This publication introduces the rationale and key elements of microdata linking and suggests new angles for economic analysis and support for policy making. It also deals with the benefits and limitations of microdata linking and the main lessons learned so far.

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

Publications Office
of the European Union

ISBN 978-92-76-13178-6