STATBEL
Belgium in figures

Microdata for statistical and scientific research,

the possibilities for researchers to obtain individual data held by Statbel

Erik Meersseman and Patrick Lusyne

INTRODUCTION

Statbel, the Belgian statistical office, collects, produces and disseminates reliable and relevant figures on the Belgian economy, society and territory. In order to produce these statistics, Statbel collects data. To do this, Statbel requests existing data from public or private institutions and organises surveys among citizens or enterprises.

The dissemination of figures is done primarily through global and anonymous statistics. This refers to results where a figure always relates to several citizens or enterprises and therefore no confidential information is disclosed. For some researchers, however, such global and anonymous statistics are not enough to carry out their research. The legislation foresees the possibility for Statbel to supply individual data (microdata) in certain cases, where each line in the file contains data from one person, family or business.

This article describes the possibilities for researchers to obtain individual data held by Statbel for statistical and scientific purposes.

LEGAL FRAMEWORK

The processing of data by Statbel and also the sharing of these data with third parties is regulated by the law of 4 July 1962 on public statistics (hereafter referred to as the "law on public statistics"). The starting point of the law on public statistics is that all data obtained and processed by Statbel for the production of statistics, is protected by statistical confidentiality.

Furthermore, as far as personal data is concerned, the *General Data Protection Regulation* (Regulation (EU) No 2016/679 of the European Parliament and of the Council of 27 April 2016 - hereafter referred to as "GDPR") and the *law of 30 July 2018 on the protection of natural persons with regard to the processing of personal data* (hereafter referred to as the "GDPR Framework Act") also apply.

Any supply of individual data will happen within the limits permitted by the legislation mentioned above. Note that the law on public statistics also protects data relating to enterprises or individuals who have died, which is outside the scope of the GDPR.

FOR WHAT PURPOSES CAN MICRODATA BE REQUESTED?

Researchers can request microdata if the data is necessary for their statistical or scientific research.

Necessary means that the research is not possible on the basis of global and anonymous data that Statbel disseminates without restrictions via the website https://statbel.fgov.be/en:

- **Be.STAT**[1], Statbel's electronic database, allows researchers to create customised tables;
- Through **Open data**[2], Statbel makes aggregated data available at a fairly detailed level.

The purpose for which the data is requested is limited to statistical and scientific research. However, this goal may be interpreted broadly. Policy preparatory research or research by private institutions are also eligible, provided that the processing happens independently, transparently and by making use of scientific methods. The results of the research should also be made public.

The result of the processing must consist of global and anonymous statistics and research reports that may not have any individual impact on the citizens or enterprises concerned. Of course, these published global and anonymous results can be freely used afterwards for non-statistical or non-scientific purposes.

[1] https://bestat.statbel.fgov.be/bestat/index.xhtml
[2] https://statbel.fgov.be/en/open-data

WHO CAN REQUEST DATA?

The law on public statistics lists the recipients who are eligible to receive microdata from Statbel. These are primarily federal, regional, provincial and local administrations and public institutions. These include the regional statistical authorities, the Federal Planning Bureau and the Statistical Department of the National Bank of Belgium.

Natural and legal persons pursuing a scientific goal are also eligible. More concrete, it concerns domestic and foreign universities, study departments and international organisations. For natural persons, it is virtually impossible to meet the security requirements (see below) imposed by Statbel to obtain microdata. As a result, (doctoral) students are only eligible if the application comes from the university itself and it also act as a guarantor for the protection of the data.

WHAT DATA IS ELIGIBLE?

Both existing (administrative) data that Statbel has requested and data that Statbel has collected from citizens or enterprises through surveys are eligible.

As far as existing (administrative) data is concerned, Statbel has to respect the conditions set by the Statistical Supervisory Committee, the Information Security Committee or the contract concluded with the provider of the data. It is mostly stipulated that Statbel may only make available data that it has processed. Researchers looking for raw data should contact the data supplier. Furthermore, for each variable (or for each cluster of variables) motivation must be given why the data is needed for the research. Survey data collected by Statbel itself are considered as one coherent set of variables, whereby the motivation should be done at the level of the survey as a whole.

Statbel only makes *pseudonymised data* available to researchers. This is data at an individual level (person, family or enterprise) where it is not possible for the researcher to identify the person, family or enterprise. This means, first of all, that the dataset may not contain any information such as a national register number, enterprise number, name, address, phone number, etc. that would allow a *direct identification*. However, the data must also be protected against *indirect identification*. This means that a combination of data may not lead to a single person, family or enterprise. This is evaluated on a dataset by dataset basis, but in general:

- the date of birth is converted to an age or age class;
- the postal code or municipality is replaced by the district, province, region or degree of urbanisation;
- amounts are rounded off, capped or expressed in classes or percentiles;
- NACE codes of enterprises are regrouped.

PRIVILEGED RESEARCHERS

The statistical offices of Flanders (Flemish Statistical Authority), Wallonia (L'Institut wallon de l'évaluation, de la prospective et de la statistique) and Brussels (Brussels Institute for Statistics and Analysis), the statistical departments of the National Bank of Belgium, the Federal Planning Bureau and the Price Observatory of the FPS Economy may, from a legal perspective, also receive *non-pseudonymised* microdata as part of their mission within the National Accounts Institute (NAI). It makes them an exception. However, the condition is that the non-pseudonymised microdata are necessary within the framework of their statutory task of producing statistics. Note that these institutions and their staff are subject to statistical confidentiality (see below) within the framework of their statutory task, just like Statbel.

LINKING OF DATA

Linking data from Statbel with data from third parties for which the researcher has obtained authorization in the context of the research project, can be an added value for researchers. Because Statbel is legally recognised (Royal Decree of 13 June 2014) as a

trusted third party, Statbel can guarantee the linking of both data and their pseudonymisation. Statbel keeps the key of the pseudonymisation so that possibly additional data can be added at a later date.

In the case of a linking, it is important that the researcher himself does not have access to data with direct identification that would enable the researcher to find out the identity of the person, family or enterprise.

STATISTICAL CONFIDENTIALITY AND DATA LEAKS

All data at Statbel's disposal is protected by statistical confidentiality. This means that the data may not be used in any way for any purpose other than statistical and scientific research. Statbel therefore imposes some security measures when providing microdata:

- the finality is contractually defined, as well as a prohibition to further disseminate the microdata;
- the applicant should have a Data Protection Officer (DPO) and a security advisor;
- the data should be placed on a secure server, located within Europe, with access restricted to the researchers involved in the project;
- it is stipulated in a contract that the output may only consist of global and anonymous statistics.

HOW TO REQUEST MICRODATA?

Only data that Statbel possesses and that represents added value for the research can be requested. In order to check whether this condition is met, the researcher should first consult a statistician from Statbel who is familiar with the research topic and the data. During this conversation, it will be determined, based on the researcher's need, which microdata will be eligible for an application. Some not unimportant aspects here are the quality of the data, the universe, the reference periods and the moment the data is available. During this phase, it will also be checked whether the application does not involve an (increased) risk with regard to data security, in which case the DPO must be consulted first.

If Statbel can add value to the research by providing microdata and there is, at first glance, no (increased) risk, the researcher can submit a formal application using a standardised form. The form can be found on the website of Statbel (https://statbel.fgov.be/en/about-statbel/what-we-do/microdata-research). The survey consists of two parts. The first part is the *microdata application form* and covers the following aspects:

- the identity of the controller (or the legal representative);
- the purpose for which the data is requested (may only be statistical and scientific research);
- the data requested (variables, reference period, source, etc.));
- the desired storage period with motivation;
- which category of persons will have access to the microdata;
- what the output will consist of (must be limited to global and anonymous results);
- the identity of the DPO.

The second part contains a *statement regarding technical and organisational measures* to prevent a breach of statistical confidentiality or data leaks. More concrete, this second part contains:

- the identification of the security adviser;
- the internal procedure concerning data leaks;
- the enumeration of specific measures to counteract the violation of statistical confidentiality and data leaks;
- the organisation's security policy;

The completed form is the basis for the deliberation of the application by Statbel and the first part of the document should therefore be signed by the controller (or the legal representative). The second part should be signed by the security adviser, the controller or the organisation's legal representative. The *controller* is not the person who is at the head of the concrete survey for which the data will be used, but the legal or natural person as defined in Article 24 of the GDPR. More concrete, it is the person

who sets the goals, provides the resources, signs the confidentiality contract and is liable if anything goes wrong. This person is, for example, the rector or administrator of a university, the chairman of a federal administration, the secretary-general of a regional administration, etc.

DELIBERATION AND AUTHORIZATION BY STATBEL

Each application will be assessed by a multidisciplinary committee within Statbel. In addition to checking whether the legal conditions have been met, it is also checked whether the requested data is proportionate to the research objective and whether the technical and organisational measures are in proportion to the risks. If necessary, additional measures are imposed to counteract the risk of indirect identification, reuse for non-statistical or non-scientific purposes or data leakage.

Based on the advice, Statbel's DPO will draw up a formal DPO advice. Ultimately, it is the director-general who, as Statbel's legal representative, gives the approval by signing the DPO's advice. Finally, a confidentiality contract is drawn up and submitted for signature to the researcher's data controller. The deliberation by Statbel and the drafting of the confidentiality contract take two to three weeks from the receipt of the correctly filled in and signed application form. The DPO advice and the signed confidentiality contracts are published on the website of Statbel: https://statbel.fgov.be/en/about-statbel/privacy/privacy-gdpr.

SIMPLIFIED PROCEDURES

The formal procedure developed by Statbel to supply microdata is limited to the information needed to correctly assess the application. If the researcher wishes to modify an existing authorization, Statbel provides a *simplified application form*. This can only include extending the retention period, adding extra variables, adding additional reference periods and extending the research purpose. The simplification with regard to the normal procedure lies in the application form, which is less extensive.

For some research, aggregated data with *a low risk of indirect identification* are sufficient to carry out the research. Because of this small risk, the data cannot be considered as global and anonymous and a deliberation by Statbel should take place. Here, too, a highly simplified application form suffices.

MICRODATA DELIBERATIONS AND COMMUNICATIONS SINCE MAY 2018

The procedure described above for obtaining microdata is applicable since the introduction of the GDPR on 25 May 2018 and the abolition of the Statistical Supervisory Committee of the former Commission for the protection of privacy (Privacy Commission). During the years 2018, 2019 and 2020, Statbel approved 209 requests from researchers to obtain microdata, 50 of them via a simplified procedure to extend an existing authorization. Due to the prior contact between the researcher and Statbel's statistician, applications are hardly ever refused. In some cases, however, an adjustment of the application was necessary.

Type of request	2018	2019	2020	Total
Regular requests	24	39	36	99
Applications by another statistical authority that is a member of the ISI	5	26	16	47
Applications within the framework of the NAI	4	3	6	13
Extending an existing contract	12	15	23	50
Total	45	83	81	209

PRACTICE – USER CASES

The data-technical evolution of the last decade has led to a thorough reform of regular statistical production within Statbel, including a much greater emphasis on secondary sources, such as administrative data. This has yielded large efficiency gains, which have been leveraged to develop increasingly advanced data products to meet the growing needs of our stakeholders (including policy departments, the academic world, etc.).

Statbel's data products span many substantive research and policy domains (economic, social and demographic data) and the innovations take different forms:

- some new data products are direct extensions of the regular data and statistics production. This is often done at the request of important institutional players. Ad hoc modules in surveys are an example of this.
- to some extent, administrative data is still undervalorized. Statbel has developed several data products that open up the sources at its disposal much better and lead to richer valorisations, making it possible to shed light on economic and social phenomena that were difficult to map until recently. The recent development of origin data and also longitudinal data series based on the National register are examples of this.
- the linking between administrative data can provide insights that remain completely obscure when research or policy preparatory work is based on a single source. The administrative census is an example of a standard product where this principle is explicitly put into practice. However, the very rich data ecosystem of Statbel allows a much freer integration of information from different databases. Moreover, it is not only about data collected by Statbel, but also about cooperation with other data holders, such as the Crossroads Bank for Social Security (CBSS) and the Intermutualistic Agency (IMA). All this is done strictly within the lines of the law.

Certainly with regard to the latter, but in fact globally, Statbel offers customisation, although large data consolidations are now arising as various customised databases are distributed to a wide variety of users. This in turn has a harmonising effect and even leads to research consortia organising themselves somewhat around our data.

The following non-exhaustive list includes a short anthology of different, rather recent research projects that make use of microdata provided by Statbel:

- **Immilab** is a project sponsored by the Belgian Science Policy Office (BELSPO) that investigates the position of people with a migrant background on the Belgian labour market. Research teams from the UA, ULB and UMons participate in this project. Statbel contributes through an innovative linking between the survey on structure and distribution of earnings (SES) and demographic data;
- **Re-invest** is a project that focuses on the link between social protection, housing and health. This project uses the longitudinal data 2004-2018 of the Statistics on Income and Living Conditions (SILC) as a backbone, enriched with Census data (2011) and with demographic information (e.g. on relocation movements). Partners in this project are the KU Leuven - HIVA, the Combat Poverty Service and the UCL.
- **SUSPENS** is a BELSPO-financed project carried out by researchers of the University of Antwerp, the ULB and the Federal Planning Bureau. The Statbel surveys SILC and Household Budget Survey are used as sources, so that the link between income, consumer behaviour and greenhouse gas emissions can be made at household level.
- The **VMM** (Flanders Environment Agency) calculates an indicator to carry out the affordability test of the integral water bill per year. Expenditure on the integral water bill is compared with the household income to provide policy advice on the affordability of drinking water. For this purpose, Statbel linked information from SILC to data from the water bill of the VMM. To protect the data from identification by the VMM, additional technical and organisational measures were imposed.
- **Causineq** and the **mortality databases** during the **COVID-19** epidemic: Causineq is a large-scale study of differential mortality according to social gradient in which the UCL and the VUB participate. From the beginning of the Covid-19 crisis, Statbel produced similar weekly mortality data that were used by various institutions and universities to map out excess mortality (Sciensano, UCL, VUB, UHASSELT).

Perhaps even more striking is that, parallel to the technical-methodological progress made in recent years, a cultural shift also took place within Statbel. The shift from a mere "statistics-producing" organisation to a major data producer has gone hand in hand with the adoption of a more open way of working and, above all, a significant increase in services to stakeholders.

CONCLUSION

Statbel has high quality data and the legal mandate to make these data available to researchers for statistical and scientific research. In addition, Statbel has the necessary expertise to create technically advanced customised data products. As a result, Statbel has evolved in recent years into an organisation that actively helps researchers by making data available to meet their concrete needs.