Distr. GENERAL

22 October 2012

WP. 2
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Seminar on New Frontiers for Statistical Data Collection
(Geneva, Switzerland, 31 October-2 November 2012)

Topic (i): New data sources

USE OF COMMERCIAL DATA IN THE GERMAN BUSINESS REGISTER

Contributed Paper

Prepared by Dr. Susanne Maus and Roland Sturm, Federal Statistical Office, Germany

I. Introduction

1. The German business register is a regularly updated database, which includes economically relevant enterprises, local units and legal units. Among the latter – introduced by the current EU regulation on business registers – are legal units situated in Germany that are identified as part of an enterprise group. Indicators for determining economic relevance are turnover, number of employees and direct and indirect holdings of economic units. The business register is used to support statistical surveys, which helps to ease the response burden on businesses. It is also itself an instrument for economic analysis, e.g. about structure and demography of the economy. The business register contains information for linking files of statistical data from different sources. Besides basic information on name and address of a legal unit, the business register provides the size of a legal unit according to its turnover and employment, economic activity according to the NACE classification, legal form and characteristics about its position within an enterprise group structure if the legal unit belongs to an enterprise group.

2. The most common and accessible sources used for the maintenance of the register include on the one hand files of administrative bodies such as the Federal Labour Agency or financial authorities, and on the other hand data from different statistical surveys of industry, trade, the services sector etc.. None of these sources provides information on enterprise group features that can be processed in automated ways. Therefore the data collection of information on the enterprise group structures is a specific issue in the processing of the statistical business register. A commercial data provider delivers updated information on legal units belonging to enterprise groups to the Federal Statistical Office once a year.

3. The paper outlines the German experience with this new kind of data source for the German statistical business register and covers aspects of data description, the data acquisition process, the production cycle and the experience concerning data quality.

II. Data description

A. Data needs

4. Due to the globalization process, complex enterprises arise, with national and international relevance. Legal units, if they are part of a complex enterprise or even of an enterprise group, do not longer individually have the competence to make all economic decisions themselves. Instead, a global decision centre undertakes this task for the entire group. Another point is that some economic data is only meaningful and available at group or subgroup level, but not at the level of a legal unit.

5. The business press prominently reports about big economic groups, but the complete formation of the units behind the reports often remains unclear. For analyzing the economic structures or consulting political decision makers, it is important to see the "whole elephant" and to have information on the links of control between legal units in order to define enterprise groups.

6. If enterprise group data was covered by the business register, a reduction of statistical reporting burden may be envisaged if only one unit of the group could be surveyed, instead of every single legal unit of the entire enterprise group.

7. With enforcement of an EU regulation[1], all EU member states have to cover information on enterprise groups within their national business register.

B. Data requirements

8. According to the legislation the national business register shall contain all resident enterprise groups as well as national truncated parts of multinational enterprise groups. The underlying definitions of the mentioned enterprise groups are:

   (a) An all resident enterprise group contains of at least two legal units, which are all located in the same country.

   (b) A multinational enterprise group contains of at least two legal units located in different countries.

   (c) A truncated enterprise group is defined as the part of the multinational group with all those legal units resident in the same country.

9. The legislation defines the characteristics, which have to be collected for the described populations above. Beyond basic information like the name, also economic and demographical figures, mainly NACE 2-digit level, number of employees, date of commencement and cessation, are obligatory characteristics for enterprise groups.

10. According to article 3 of the legislation, enterprise groups can be identified through the links of control between their legal units. The underlying statistical concept of control can be reduced on four main criteria used in practice to determine the link of control between two units:

   (a) A legal unit directly holds more than 50% of all the shares of another legal unit (direct control).

   (b) A legal unit indirectly holds more than 50% of all the shares of another legal unit by holding shares of subsidiary enterprises (indirect control and indirect cumulative control).

   (c) The accounting system of a legal unit is fully integrated into the system of another legal unit (consolidated accounting system).

[1] (EC) No 177/2008 of the European parliament and of the council of 20 February 2008 establishing a common framework for business registers for statistical purposes and repealing Council Regulation (EEC) No 2186/93.

(d) Administrative sources report the link of control even for a shareholding of less than 50%.

11. Background of the concept is an influence of a parent enterprise on the strategic decisions of a subsidiary enterprise in the medium term. For a more detailed overview see Sturm, Tümmler and Opfermann (2009)[2].

C. Data setting

12. The information purchased from the commercial data provider is split in two data sets. The first file includes all German legal units which belong to an enterprise group, their all their shareholders located in Germany as well as the direct parent company and the global group head if located abroad. New to the population of statistical units in the business register is the coverage of natural persons as group heads or minority shareholders.

13. The second file contains relations of ownership and links of control between two units. According to the statistical concept of control, the group structure can be identified. Minority shareholders are included in those cases where the shares belong to a unit which is part of an enterprise group.

III. Data acquisition process

A. Considerations for data purchase

14. The most important data sources for the German business register are administrative sources. Unfortunately, the information if and how a legal unit belongs to an enterprise group is not contained in the so called electronic trade register, so is not available in a technically operable and efficient way for the purpose of the national register so far.

15. A statistical survey to collect information on enterprise group features would – in principle – be a way to collect enterprise group information. But according to the general strategy of official statistics in Germany to reduce reporting burden, establishing a new statistical survey does not seem to be a promising strategy.

16. Statistically relevant data is also stored and offered by commercial data providers. As in other European countries, private data providers collect – to large extent in manual work – such information from different sources, mainly based on systematic analyses of the mentioned trade register. The workload is enormous and only profitable if there are several circles of clientele. In the case of enterprise group data for the German business register, it was decided to gather first experiences with this way of data acquisition. Since 2005 the Statistical Offices in Germany buy the relevant information on enterprise groups from a commercial data provider every year.

B. Call for tender and main criteria for the choice for a provider

17. The acquisition of the data from commercial providers obeys the purchasing management requirements of the public sector in Germany. According to the expected financial volume, a Europe-wide call for tender was launched.

18. Main subject of the call for tender is the description of the specifications of services including notes regarding an offer, minimum requirements, additional and optional requirements. The purpose of the notes mainly is to inform the potential provider about the structure of the offer and about the consequences in case of acceptance of the tender. The minimum requirements include aspects of data quality, like completeness and representativeness, doublets, core variables and

[2] Sturm, Tümmler and Opfermann (2009): Unternehmensverflechtungen im statistischen Unternehmensregister, Wirtschaft und Statistik 8/2009.

appropriate identification numbers, as well as further conditions, for example time stamps, basic support or illustration of prices. Additional requirements concern issues like the range of further identification numbers, additional variables or extensive support.

19. A number of offers were received, which considerably differed regarding appropriateness. The decision procedure is based on a public catalogue of criteria, which includes all aspects of the specification of services. The criteria are categorized in criteria, which 1. lead to the exclusion of an offer, 2. should be fulfilled and 3. are 'nice-to-have'. All incoming offers are evaluated by a group of experts according to the catalogue of criteria, which are weighted according to the importance of the specific aspect. The commercial data provider with the highest scored offer gets the acceptance of the tender.

IV. Production cycle

A. Interaction with the data provider

20. Each cycle starts in midyear with a request for delivery of the yearly data material from the commercial data provider. Possible adjustments on the material are discussed and introduced during the late summer. The delivery of the data is scheduled for end of October. The first step of data processing is the validation process, whose results are reported to the data provider. In a second step, incorrect data has to be reviewed and corrected by the data provider. Afterwards, either a corrected data set or just corrected parts, if the correction of the errors is easy to implement, have to be delivered. This new material is then tested during the validation process once more. In theory, the data provider is requested for corrections, as long as incorrect data exists.

21. The processing of data correction and therefore the quality of data also depends on the timeline of the business register cycle, in which the processing of enterprise group data forms one of many elements of the yearly production process. By beginning of the next year, the enterprise group data processing within the business register starts. All invalid data found later, which should not be 'fundamental' anymore, are reported to and discussed with the data provider, but are only corrected and implemented for the next purchase.

22. Based on experiences from the previous years, the data provider delivers data maximum two or three times for one reference year. Due to the processing of the data in the business register from January on, the buy-off of the material takes place at that time as well. For questions concerning the material or background of the data acquisition, a contact person is available the whole year.

B. Matching with business register units

23. After validation and the buy-off of the data, the information about enterprise groups has to be added to the units in the business register. This is done by linking each single legal unit from the database of the commercial data provider to the corresponding legal unit of the main population in the business register with help of numerical identifiers and address comparison. Not included in the linking procedure are branches, natural persons and foreign units, since those are not stored in the main business register population. Information about this population is kept in the enterprise group data base beyond the business register.

24. As some variables are delivered in a different way as they are included in the business register, previous adjustments are necessary. A major prerequisite and effort is the generation of the so-called trade register number, which is a composite number of the identification code of the local register court, the kind of the register an enterprise is registered and the identification number the register court has given to the enterprise. The commercial data provider delivers the postal code of the local register court instead of the identification code. So the transformation of the postal code to the identification code of the local register court has to be done. The resulting register number is the most important identifier for linking the enterprise group data set with the national business register. If the linking procedure on the basis of the trade register number is not

successful and no corresponding legal unit can be found, the matching must be done manually by comparing the addresses of both data sets. Units matched manually in previous cycles can be linked on the identification number of the data provider, or in case of wrong or insufficient information of the trade register number.

25. For all units, which can be matched, information of three variables from the enterprise group material are stored within the legal unit in the business register: the identification number of the data provider for each unit, the one of the group head and the information on the position of the legal unit in the hierarchy of the enterprise group.

V. Experience on data quality

26. To get an assessment of the quality, first of all it is necessary to know the volume of the delivered data. The following table shows the total number of units in the data set for legal units with the reference year 2010. It can be broken down to sub-totals which have different relevance regarding the processing with the core business register. Although the business register certainly contains branches, they are currently not processed with respect to enterprise groups, because local units are of minor relevance for the purpose of enterprise groups. The information on natural persons is not matched with the core register as well, since natural persons are not included in the register as long as they do not run a non-incorporated firm. Besides, information on natural persons is delivered in poor quality; often only the name is available. Nevertheless natural persons are very important for the description of the whole group structure.

Table 1: Sub-totals of the commercial data on legal units

Sub-totals	**Number of units (reference year 2010)**
1. Total data set	955.503
2. Legal units without branches	822.387
3. German legal units without branches	776.239
4. German legal units without branches and without natural persons	**570.995**
5. Enterprise groups (based on the sub-total in the 2nd line)	207.480

27. As the business register only includes legal units resident in Germany and due to the exclusion of branches and natural persons, about 60% of the delivered units are subject to the processing with the core business register. After the matching procedure 469.806 units have a corresponding unit in the German business register, of which 65.582 units have been ceased or liquidated. That means about 18% of all the potential units in the enterprise group material have no corresponding unit in the business register.

28. One reason could be the differences in the coverage. The business register only contains units with a yearly turnover of at least 17.500 Euro and/or at least one employee, who is subject to social insurance contribution, while the database of the private data provider includes all enterprises registered in the trade register having a certain score of credit rating. A second reason for not finding a corresponding unit in the business register is partly poor quality of the concerned unit in the enterprise group data like existence of doublets for example due to consistency errors. Another reason could be the different time stamp of the business register compared to the data sources the data provider generates the historical material from.

A. Measures of quality checks

29. Before the matching procedure can start, it is necessary to validate the data of the commercial data provider. The validation tests are necessary to identify missing and implausible values within the material, as well as compared to materials of previous years. The implementation of new and adjusted tests thereby will not be finished due to a development of the database and also due to a development of the use of the data. Tests are also implemented according to the feedback of the users of the enterprise group data.

30. For both data files (described in II.C.) several tests are carried out, which are categorized into three levels of error classes. The single cases of all of those classes are reported to the data provider. While the cases of the first error class have to be corrected immediately, the cases of the second category can be corrected at the next purchase. The values of the cases in the third error class are implausible, but might be true. A comment on these cases by the data provider is sufficient for further processing.

31. The first error class for legal units includes tests concerning doublets, incorrect or incomplete identification numbers and wrong information of names and addresses. The second category includes cases with errors which do not lead to stop the whole processing of the data. Mainly implausible values of more important variables are categorized in this class. The last category includes tests on implausible reference dates and implausible values of variables, which are of minor importance, because the information can be generated from another variable as well or is not relevant for the processing of the enterprise group data. For example a wrong legal form means only minor trouble as the information is already available in the business register. Another example are incorrect values in an identification number for German regions. While the needed information for the mapping of the cases to the German Bundeslaender can be taken from the postal code as well, it is less important to have the regional identification number corrected by the commercial data provider.

32. The first error class for the validation of the link dataset includes tests mainly concerning different information in the variables of the link dataset compared to the dataset containing the legal units. Currently no tests of the link data are categorized in a second error class. The third category has tests on incorrect and missing redundant information (e.g. shares in percentage differ from the information given in the variable “share category”).

33. For the dataset of the legal units, 52 validation tests are implemented. 32 of those produce cases of the first error class, 7 of the second and 13 tests uncover errors of the third error class. For the dataset with the link information, 27 tests are implemented in total. 20 tests for the first category, 0 for the second and 7 in the least problematical category. The validation procedure is in progress and comprehensive tests are added to each cycle. For the merge procedure with the business register mainly the dataset with the legal units is relevant. This is the reason for the higher number of tests for this material at the moment.

B. Results of plausibility and quality checks

34. The following table lists the number of cases in the corresponding error classes detected during the validation procedure of the two datasets. Although the first category includes the highest number of tests, only 16% of all errors in the dataset including the legal units arise in error class one. A different picture comes up for the validation of the dataset including the links of the legal units. 99% of all errors are of category one. This result might be an indication for quality differences between the two datasets; the better observed dataset seems to have comparably less serious errors.

35. The number of errors in the error class one of the legal unit file is mainly concentrated on four tests. Most errors, at least 10.000, occur when testing the postal code of local register court and legal units, when testing the trade register number and when testing the address of an enterprise. The identified errors in the link file are concentrated on the test on similarity of the legal form information compared to the legal units file.

Table 2: Error classes

Reference year 2010	Dataset with information on:	
	Legal units	Links of legal units
Number of legal units and links in total	955.503	2.107.794
Error class one	194.568	97.545
Error class two	3.244	0
Error class three	1.032.531	775
Cases including errors in total*	1.230.343	98.320

* A legal unit or a link can have more than one error.

36. The errors are distributed on 591.306 units in the data set with the information on the legal units and on 98.237 units in the data set with the link information. Due to much less tests for the link data set, the error rate is only 5% and therefore much less compared to the one for the data set with the information on legal units: 62% of all units fail at least one of the tests. Units containing errors of category one would be excluded from further processing if the erroneous information is not corrected. These serious errors are distributed on only 73.787 units in the data set with the information on the legal units and on 97.530 units in the data set with the link information. The total error rate for first class errors is therefore only 8% in the data set with the information on legal units and 5% in the material with the link information.

37. One of the most important information for analyzing enterprise group structures is the one on the position of a legal unit within the enterprise group structure. The following table shows the frequencies of the position of German legal units, which are part of an enterprise group. For 109.733 units the information on the position is not available from the data set with information on legal units, this is about 14% of all German legal units.

Table 3: Group position

	Number of units*	In %
(German) group head	180.021	27,0
Intermediate	48.937	7,3
Pure subsidiary	437.548	65,6
Total	666.506	100

* Number of German legal units without branches.

38. The quality of data can be measured by the fillings of the key variables. The matching procedure is more difficult and could possibly fail in total, when basic information about an enterprise is incorrect or not delivered. Ideally the matching of two units is done by a numerical unique identification number. For new arising units, which may be included in the material, but are not yet included in the business register, or for units without correct information on identification numbers, the matching has to be done by using the information of the name and the address of an enterprise. The following table contains the absolute and percental figures of cells with filled information of key variables, compared to all German legal units with a potentially corresponding unit in the business register.

Table 4: Cell filling of key variables

Key variables	Number of units* with cell information	In % of German legal units without branches and natural persons
Identification number of the group head	550.843	96,5
Trade register number	541.764	94,9
Name	570.995	100,0
Street	563.025	98,6
City	567.142	99,3
Postal code	564.813	98,9
Postal code of register court	482.621	84,5
Value added tax number	132.058	23,1
Identification number of the private data provider Duns and Bradstreet	577	0,1
Regional code	450.016	78,8

* Number of German legal units without branches and natural persons.

39. All in all, the quality of the information, which is necessary to match the units with the corresponding ones in the business register, is rather complete. Regarding most of the additional economic information, the picture is quite different: While for 94% of all German legal units, which are processed, the economic activity code (NACE) is given, for less than the half of the units (48%) information about turnover is available and for only 30% of the units information on the number of employees.

C. Corrections of data by data provider

40. The results of the validation process, which are presented in this section of the paper, refer to the validation procedure for the final delivery of the data material with reference year 2010. As described before, the data provider should at least correct errors of error class 1. The correction process reduced the numbers of errors compared to the first material about 42% for the data set of legal units and about 2% for the data with the link information.

D. Discrepancies between administrative and commercial data

41. For each unit that can be successfully linked to a register unit, further quality checks on the economic variables can be done. The next table shows the differences between the information for the turnover and the number of employees of commercial data, compared to the administrative data stored in the business register. Only for 29% of all linked units information on turnover is available from the commercial data source, for information on the number of employees 42% of all matches. The small rate of economic information results not only from missing values in the commercial data, but also from different timestamps of the information in the commercial sources.

42. The turnover and employment figures received by the business register from administrative sources serve as reference, when evaluating the quality of the figures received from the commercial source. For both variables a clear overestimation in the commercial data can be detected. The sum of total turnover of all units successfully linked to the business register referring to the relevant information in the commercial data results in 197% of the sum regarding to the business register and 166% of the number of employees. Similar results have been found for previous cycles. However, a reduction of the differences can be noticed over the years. For further analyses on previous cycles see Kleber, Sturm and Tümmler (2010)[3].

[3] Kleber, Sturm and Tümmler (2010): Ergebnisse zu Unternehmensgruppen aus dem Unternehmensregister, in Wirtschaft und Statistik 6/2010.

Table 5: Deviation of economic variables (absolute)

	Business register data	Commercial data	Over-estimation	Number of enterprises
Turnover (in bn. Euro)	2.432	4.783	97%	116.370
Employees (in m.)	11,6	19,2	66%	168.792

43. The following table gives an impression of the distribution of the differences. The differences of information for all linked units including the relevant information are categorized into five groups. An example of reading the table is that the information on turnover is the same in the two sources, commercial and administrative for only 0,5% or 597 enterprises of all linked units, which include the relevant information. All in all, 37,7% of the linked units have an absolute difference of 10% at maximum compared to the administrative data in the business register, but 34% of the linked units show differences between 10 till equal 50% of value and 29% of the units differ by more than 50% of value.

44. The economic relevance of enterprise groups would be extremely overestimated if taking the information of the commercial data provider fully and uncritically into account. Only a smaller fraction of the obvious differences in the economic information can be explained by different underlying concepts. While, as described earlier, the business register includes only enterprises with employees which are subject to social insurance contribution, the database of the private data provider includes all enterprises registered in the trade register having a certain score of credit rating. A great extent of the differences may be caused by multiple processing of different sources the commercial data relies on. So in many cases the information are on the one hand taken from annual reports, which often publish data referring to the whole enterprise group, and bearing the risk to be wrongly attributed to parts of the group by the commercial data provider. On the other hand single units of the specific group could in addition deliver information on that single unit in the frame of surveys. In the summing up, the information of the commercial data could be overestimated due to exaggerated information for a single legal unit and double counting of values.

Table 6: Deviation of economic variables (categorical)

Deviation in categories	Turnover		Employees	
	Number of enterprises	In % of linked units including the information	Number of enterprises	In % of linked units including the information
Equal to 0 %	597	0,5	19.018	11,3
>0% till equal 5%	29.168	25,1	12.432	7,4
>5% till equal 10%	12.869	11,1	16.018	9,5
>10% till equal 50%	40.030	34,4	64.620	38,3
More than 50%	33.706	29,0	56.704	33,6
Missing value	287.854	71,2*	235.432	58,2*

* In % of all linked units

45. The last comparison of information between commercial and administrative data that we have done in this study, is the comparison of the information for the economic activity. For 393.360 units the information is included in both sources and the analysis can be done. The following table shows the results. For about one third of the linked units, the information is identical in the deepest categorization of the economic activity, the German subdivision of NACE[4] code at 5 digit level. But for about the same fraction of linked units (34,1%), even the 2-digit NACE code does not fit in both sources.

46. A possible reason for the differences might be the one described before for the turnover and employee information: As most of the information is taken from a variety of sources and is not

[4] Nomenclature statistique des activités économiques dans la Communautés Européene.

directly surveyed from the enterprises belonging to a group (as administrative data does), the information of the commercial data could refer to the whole group, instead of the specific legal unit of the group as it is in the business register.

Table 7: Matches of the information on the economic branch

	Matches	In %
National NACE (5-digit) sub-division	143.710	36,5
NACE (4-digit)	194.630	49,5
NACE (2-digit)	257.375	65,4

47. The conclusion of quality analysis of economic variables is, that information on economic issues taken from administrative sources can not be questioned or substituted by data from the commercial source. Information on turnover and the number of employees of commercial data is not used by official statistics.

VI. Summary

48. The paper outlines the process of data purchase of enterprise group information for the German business register and the experiences on the data quality of commercial data sources. After describing the call for tender procedure, quality aspects of the commercial data are discussed. The main results of the paper are:

(a) Need for enterprise group data by law

(b) Preference for commercial data against installation of new surveys or use of administrative sources

(c) Validation procedure of the commercial data extensive, but still in progress

(d) Validation allows sensible use of the commercial data for the relevant information in official statistics (group structures)

(e) No use of the commercial data for further economic analyses in official statistics.

49. The development of validation procedures allows for growing insight in the usefulness and the limitations of the new data source. A clearer assessment of the potential and the perspective of using commercial data in future more extensively for the purpose of collecting information on group structures have been gained. Nevertheless the validation procedure is not finalised so far and has to be differentiated and deepened for sub-groups of the commercial data.