The integration of enterprise data

Gürke, Christopher
Research Data Centre of the Federal Statistical Office Germany
Gustav-Stresemann-Ring 11
65189 Wiesbaden, Germany
E-mail:christopher.guerke@destatis.de

Zwick, Markus
Research Data Centre of the Federal Statistical Office Germany
Gustav-Stresemann-Ring 11
65189 Wiesbaden, Germany
E-mail:markus.zwick@destatis.de

The Research Data Centre (RDC) in Germany offers a wide range of possibilities to use official micro data sets. The researchers can apply for so-called 'scientific use files', which are anonymised micro data usable outside of the statistical office. But anonymisation is always the reduction of information. This is the reason why the RDC provides access to more detailed data inside the statistical office. The next step of this kind of data access is to develop a real 'remote access' application. Germany and also Eurostat, with Germany as an important partner, work in different projects on this future technology.
Data access is one task of the RDC in Germany. Another task is to develop new data sources for the scientific community. One project is the matching of different business data sets. The core of this matching project is the German business register. On the one hand the new combined data sets include much more information then the 'stand alone' data sets. On the other hand this is a great problem in the process of anonymisation. Remote execution and in hopefully near future remote access are the only way for researcher to get access to these complex combinated data sets.
The paper gives an overview of the different ways of access to intergrated business micro data sets from official statistics in Germany.

1. Introduction

Since the research data centres (RDCs) of the statistical offices in Germany were set up in 2001, they have become firmly established and today empirical science is unimaginable without them.
Access to microdata of official statistics in Germany is possible via Scientific Use Files (SUF), Public Use Files (PUF), via data laboratory[1] and remote execution through the Research Data Centres (RDC) of the national statistical offices (NSO).
The paper is organized as follows. Section 2 reviews the current micro data access in the research data centres of the NSOs. Section 3 provides information about the the projects "Official company data for Germany" (AFiD), "Integrated Enterprise Data for Germany" (KombiFiD) and "An

[1] A data laboratory is a secured room in a statistical office especially designed for researchers granted with right of access to microdata. Such a room is equipped with special features preventing the transmission of any kind of information to the outside world.

informational infrastructure for the E-Science Age – On the way to remote data access for business data" (InfinitE). Section 4 offers concluding remarks.

2. Current micro data access

Access to selected official German microdata is provided in four different forms. Researchers are able to use the microdata by

- Public Use Files (PUF),
- Scientific Use Files (SUF), also called Micro data files Under Contract (MUC),
- Remote execution,

and/or

- Data laboratories.

These forms differ amongst others in their level of anonymity. The levels of anonymisation are absolutely, de facto and formally anonymised data. The next sections informs about the relation between the provided products and the grade of anonymisation.

2.1 Public Use Files

As absolutely anonymised microdata, standardised PUFs are available to all those who are interested both within the country and abroad. Due to anonymisation, PUFs contain only selected variables. As a rule, variables with a high degree of subject-related detail are aggregated. In most cases, detailed regional breakdowns cannot be made on the basis of PUFS. The reduction of information leads to lower demand of PUFs in research analysis.

Therefore the RDCs of the NSOs provide so called CAMPUS files. This kind of micro data does not primarily aim to retain as much of the analysis potential of the data as possible. The data are intended for training application-related statistical methods with complex sets of official micro data at universities. The characteristics and the sample concept are selected in such a manner that specific topical areas can be processed in teaching in sufficient precision.

2.2 Scientific Use Files

In the form of de facto anonymised microdata, the research data centres offer the microdata of common statistics as standardised SUFs for off-site use to users from the scientific community. These data have a far greater information potential than Public Use Files and they are well-suited for large part of the scientific data analyses. An off-site use is possible at research institutions which are governed by German law.

2.3 Remote execution

In Germany remote execution is the only form of access permitting the analysis of official "original"[2] micro data. In this concept, the data user and the data producer communicate by program syntax. Therefore the data user does not have direct access to the data. The user gets a

[2] Original means that only direct identifiers like name, adress, etc. are excluded from the micro data file.

structure file, this is a absolute anonymous micro data file which contains the variables and characteristics of the original data so that the researcher can develop their analysis programs. Inside the RDC, the staff applies the programs on the original data and checks the output afterwards on anonymity.

2.4 Data laboratories

There are PC workplaces at all 19 locations of research data centres where de facto anonymised microdata can be analysed by domestic and foreign guest scientists on the protected premises of the statistical agencies. Here de facto anonymity is achieved not only by an anonymisation of the data (as in the case of Scientific Use Files) but in combination with a controlled data access. This is why these data may contain much more detailed information than the Scientific Use Files submitted in the form of data files.

The kind of data use listed above may also be combined with each other. For instance, parts of a data record analysis may take place at a workplace for guest scientists, while other analyses of that survey may be performed via remote execution. Such combination often makes sense especially with longer-term research projects.

It is obvious that in all products of data access the employees of the RDCs and the users of official micro data files are confronted with a high amount of workload. Especially for the remote execution and the data laboratories the effort rises with every data user. Equally the more complex micro datasets are integrated as official statistic products the more expensive the access to micro data with respect to labour costs will get. The next section will inform about three projects. The first two deal with complex longitudinal data. The third one addresses the development of new strategies for data access e.g. remote access.

3. Projects

Among the provision of micro data access the research data centres of the statistical offices in Germany conduct several national and international projects in the fields “implementation of new micro data” and “modernisation of the existing infrastructure”.

The following sections describe three projects which deal on the one hand with matching procedures of enterprise-level data and on the other hand with remote access.

3.1 AFiD

In the past business micro data were offered only as longitudinal micro data on the level of enterprises for scientific analyses. Due to the growing number of enquiries for temporal cross-sectional data, the research data centre of the Statistical Offices of the Federal States launches the project „Amtliche Firmendaten für Deutschland“ (AFiD). The aim of the project AFiD is to merge official business micro data over time and different fields of official statistical business micro data. The underlying idea of the panel data is to advance the analytical potential and to offer information about enterprises and their local units with respect to different official statistics, different times and

different thematically issues. Nevertheless for matching data of different official statistics a unique identifier has to be used like the business register. The German business register is a regularly updated database of enterprises and local units with a taxable turnover from deliveries and output and employees subject to social insurance contributions. Evaluations of business register data on the number of enterprises and local units and their employees, who are subject to social insurance contributions, and their sales (turnover) reveal economic structures in Germany.

Besides the business data the AFiD project group developed also different modules that contain environmental information and information about earnings. These modules are longitudinal micro data sets which can be used additive to selected AFiDpanel datasets and offer a wide range of specialised scientific analyses.

3.2 KombiFiD

For many economic analyses longitudinal micro-data on the level of enterprises is needed. In order to answer certain research questions it is also often necessary to integrate enterprise-level data from different sources. In Germany, the access of the scientific community to such integrated longitudinal micro-data has improved considerably during the last ten years, due to the work of the research data centres of the Federal Statistical Office and the statistical offices of the Länder, the research data centre of the Federal Employment Agency and the research centre of the Deutsche Bundesbank.

One important step that has not been taken yet is an integration of enterprise-level data across the borders of different german data producers. This is where the project "Integrated Enterprise Data for Germany" (KombiFiD) comes into play: The purpose of this project is to merge enterprise data of the German Federal Statistical Office, the Land Statistical Offices, the Institute for Employment Research (IAB) and the Deutsche Bundesbank. Besides, a modification of the present legal requirements is aspired in the long run, in order to facilitate the combination of data sets in the future.

Through the process of data integration carried out in the KombiFiD project new possibilities for economic research will be opened up and it will be possible to analyse economic processes in a more detailed and comprehensive way. A challenge in the project context relates to the process of record linkage in the data integration process. The initial situation in the KombiFiD project is favourable insofar as there is a sufficiently broad range of overlapping variables. What's more, the variables are of high quality in all data sets. Useable are at least the name of the company, the place where the company headquarter is located and detailed information about the economic branch. In some cases it will also be possible to make use of data about the number of employees and the legal form (Rechtsform).

3.3 InfinitE

A real remote access application, which is fully automated and does not require any manual handling, is a vision for the future which has not been achieved in countries with a legal frame like Germany either. The project InfinitE is necessary to take first steps towards that goal. Furthermore

a "RDC in RDC" solution where data of one RDC can be processed in another RDC, using remote access, would be a possible next step for Germany. This can be used as a test implementation for real remote access to be established later and permits shifting activities towards more data exploration, data documentation, internationalisation and less visitor care.

Before a remote access application can completely be implemented, however, many technical, legal and – the focus of the project – methodical problems must be solved.

The development of basic strategies for producing anonymised data structure files which allow checking a program run for syntactic and semantic errors is one of the project's purposes. Therefore criteria on data structure files have been established which are necessary for a useable remote access for both sides, the researcher and the NSI staff. Current data structure files – of the bases of remote data execution, described above – allow only syntactic checking. Methods that might be applied to produce such data structure files are in particular the data perturbation methods of multiplicative stochastic noise, multidimensional microaggregation and multiple imputation. First investigations of the implementation of these methods are initiated. The project team uses a concrete scientific objective regarding labour market development to ensure, that the data structure files will comply with scientific standards.

The second purpose is the development of standardised and completely automated checking of results. By now checking the results post-tabular is always time-consuming and labour intensive. Results of remote data execution (even intermediate results) and of activities performed at safe centres are checked for confidentiality before being released. Such control is extremely difficult for complex tables, large estimation outputs and the combination of both in the same data sets. Before automatic output checking can be realised all sorts of results have to be categorised in safe and unsafe.

Together with the project partner, the Institute for Applied Economic Research (IAW), the IAB, the State Statistical Offices of Berlin-Brandenburg and Hessen, the Federal Statistical Office is willing to extend the current state of knowledge and to widen the issue and to perform a systematic comparison between data-based and result-based safeguarding of the protection of the carriers of variables with regard to the analysis potential.

4. Conclusion

Considering the examples of data access to German micro data sets progress has already been made and results have been obtained in this field. With the project »Official company data for Germany« (AFiD) the statistical offices, together with other partners, deal with highly complex data sets, which offer high analysis potential even for cross-section and panel data of economic statistics.

The feasibility study KombiFiD deals with the matching of micro data across the boundaries of different data producers. The developed micro data is available in the research data centres of the data producers since April 2011.

The project InfinitE forms an important bridge between the developments in improving data access channels for the scientific community over the last few years and the concepts planned already

today for the future by the research data centres. It is a major milestone on the way towards real remote access.

In the long term, real remote access seems to be the only feasible solution both nationally and internationally; all the more so as a method, once developed, can rapidly be transferred to other surveys and could allow "just in time" delivery of data. The technical developments have reached a phase where online access is possible from anywhere or will be possible soon with the relevant range.

REFERENCES

Bender, S., Himmelreicher, R., Zühlke, S., and Zwick, M. (2010) Access to Microdata from Official Statistics. In: Building on Progress, Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences. Vol. 1, Budrich UniPress Ltd, 215-230

Brandt, M. & Zwick, M. (2009): An informational infrastructure for the E-Science Age – On the way to remote data access for business data, conference paper "New Techniques and Technologies for Statistics", Brussels.

CENEX-ISAD (2008a) Report of WP1. State of the art on statistical methodologies for inte-gration of surveys and administrative data

CENEX-ISAD (2008b) Report of WP2. Recommendations on the use of methodologies for the integration of surveys and administrative data

Cohen, W.; Ravikumar, P.; Fienberg, S. (2003) A Comparison of String Metrics for Matching Names and Records. Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation, 73-78

Drechsler, J.; Dundler, A.; Bender, S.; Rässler, S.; Zwick, T. (2008) A new approach for dis-closure control in the IAB establishment panel - Multiple imputation for a better data ac-cess, in: Advances in Statistical Analysis, 92, 439-458

Fellegi, I.; Sunter, A. (1969) A Theory for Record Linkage, in: Journal of the American Sta-tistical Association, 64, 1183-1210

Groves, R. et al. (Eds.) (2002) Survey Nonresponse. New York: Wiley

Herzog, T.; Scheuren, F.; Winkler, W. (2007) Data Quality and Record Linkage Techniques. New York, Berlin: Springer

Monge, A.; Elkan, C. (1996) The field-matching problem: algorithm and applications, in:

Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 267-270

Schafer, J. L; Olsen, M. K. (1998) Multiple imputation for multivariate missing-data prob-lems: A data analyst's perspective; in: Multivariate Behavioral Research, 33, 545–571

Zwick, M. (2007) CAMPUS Files - Free Public Use Files for Teaching Purposes. In: Schmollers Jahrbuch : Journal of Applied Social Science Studies, 2007, vol. 127, issue 4, pages 655-668