Cross-institutional Integration of Business Data: Results of the German KombiFiD Project

Michael Konold
Research Data Centre of the Federal Statistical Office of Germany

This version: August 15, 2009

Abstract

This paper provides an overview over a German project concerned with the cross-institutional integration of enterprise data. It starts off by giving some background information on the situation in Germany and by illustrating the possibilities and challenges that arise from this situation. After that, the project and its goals are introduced. In part three, empirical results are presented. The concluding section draws comparisons to other projects and highlights some future perspectives. With respect to those, specific attention is paid to cross-institutional data integration as a way to limit the burden of respondents and as an alternative to additional surveys.

1. Introduction

In many European countries efforts are being made by official statistics to integrate business data from different sources. Sometimes such efforts aim at a combination of register data and survey data. In other cases the objective might be a linkage of survey data and process generated data or an integration of different registers. Three major reasons can be given for such activities: First of all, the matching of existing data sets occasionally constitutes a faster and cheaper way to gather certain information than the collection by means of a survey. In the second place, it can also be a possibility to reduce the reporting duties of enterprises and their establishments. Finally, the integration of data from different sources often enables novel and more comprehensive economic analyses. There are usually some methodological problems to solve though: unique identifiers – for instance – may not be available or relevant units may not be included in all data sources.

A German project directly located in this context is the project “Combined Firm Data for Germany” (*KombiFiD – Kombinierte Firmendaten für Deutschland*). It is carried out by the Federal Statistical Office, the Institute for Employment Research of the Federal Employment Agency, the German Federal Bank, the Leuphana University of Lüneburg and the University of Applied Sciences Mainz and funded by the Federal Ministry of Education and Research (BMBF). The paper introduces this project and starts off by highlighting some of features of the situation in Germany. One of these features is the fact that official business data is collected independently by several institutions (Statistical Offices, Federal Employment Agency, German Federal Bank).

A second important feature is, that due to the current legal situation cross-institutional merging of data cannot be carried out without enterprises giving their consent. The paper illustrates the challenges that arise from this constellation and focuses specifically on some issues in the context of record linkage and the relevance of certain drop-out processes. Moreover, the role of the German business register and process generated data sets is explained. In the second half of the paper, preliminary empirical results will be presented. At the moment, these are limited to one specific area. More comprehensive information will probably be available in September.

2. Integration of Business Data: The Current Situation in Germany

With respect to the collection and dissemination of official business data, the situation in Germany is characterized by the fact that there a several independent data producers. First, there are the statistical offices, the Federal Statistical Office and the statistical offices of the German *Länder*, to be more precise, which gather a broad range of data on enterprises and their establishments, trade, taxes, the labour market and other things by means of surveys and from administrative sources. The second big institution is the Federal Employment Agency, which primarily prepares process-generated data, especially on establishments and the labour market. Finally, there is the German Federal Bank, to which enterprises report on foreign direct investments and which also collects – among other things – corporate balance sheets and financial data.

All of the aforementioned data producers prepare different kinds of official reports on a regular basis and offer certain ways of micro data access for science through their Research Data Centres (RDC).[1] Hence, an access to important data sets for in depth economic analysis is possible. There is, however, one severe limitation: Micro data from different data producers cannot be combined. The reason for that is the current legal situation in Germany, which allows data merging between the Statistical Offices, the Federal Employment Agency, and the German Federal Bank only in cases where the respondents explicitly approve of this procedure. The consequences are as follows: On the one hand, certain important pieces of information (e.g. detailed information on the activities of enterprises in foreign countries/markets and information on their personnel structure, the structure of qualifications respectively) are not available for a joint analysis, which is unfortunate for science as well as managerial decisions and political institutions, which draw on the expertise of scientific studies. On the other hand, there is no possibility to use data integration as a way to further reduce enterprises' reporting duties. As certain data is collected by more than one institution, a potential for rationalization therefore remains unused.

In order to show that a cross-institutional integration of data is possible and to find out what works best, in order to evaluate the exact possibilities and challenges, and in order to pinpoint the legal changes that would be necessary to carry out the rele-

[1] Comprehensive information about the RDC can be found on http://www.forschungsdatenzentrum.de, http://fdz.iab.de and http://www.bundesbank.de/vfz. All three internet sites provide information in English. Cf. Zühlke et al. (2005), Konold (2007), Kohlmann (2005), and Lipponer (2003) for further details.

vant operations in the future, an empirical study is needed, and that is to say, that a sample of enterprises has to be asked for their consent. This is where the KombiFiD project comes in. Provided all works out as designed – two prerequisites for that are a satisfactory response by the targeted enterprises and a drop-out process which is not radically biased – the project will be able to draw the relevant conclusions. It will also be able to provide a very rich and novel data set that will be available for scientific analysis. Details of the KombiFiD project are presented in the following section.

3. The KombiFiD Project: Methods and Challenges

For the KombiFiD project, a sample of about 55.000 enterprises has been drawn. The economic branches covered are manufacturing, construction, wholesale, retail, and services. The project will therefore refer to a broad range of economic activities, which is important with respect to the possible conclusions. In order to keep things straightforward, the focus will be on the enterprise level.[2]

All of the selected enterprises were contacted in April 2009 in a postal survey and asked to give their written consent to the described combination of data within the context of the project. Enterprises with no response were reminded twice. At the beginning of August, the response rate was close to 43 percent (This comprises cases in which enterprises informed about insolvency, closure or a change of the bearer as well as cases which still need further clarifications). 25 percent of the 55.000 enterprises had given a positive response (assent). Considering that participation is voluntary and that the issue is a relatively complex one, the outcome is very good. It is also worth pointing out that there is still a significant number of replies per week (The field period will be completed by the end of August).

The next steps that have already been taken up are bias analyses and record linkage. As for the bias, it is important to ascertain if the data fulfils the basic needs of the project. It would be a problem if assent was very unevenly distributed among – for example – size classes or economic branches, or if certain categories had no enterprises with approval in it at all. Therefore, the analysis of bias related issues has been taken up early on. Preliminary descriptive results are presented in the following section.

The second – and central – area of analysis and study is of course the complex of record linkage (data matching), which in KombiFiD is carried out by the Institute for Employment Research of the Federal Employment Agency. As the statistical offices, the Federal Employment Agency, and the German Federal Bank collect data independent from one another, it is not surprising that there is no unique ID that can be used to match the relevant data right away. Instead of that, the situation is more complex. There is an identifier on the level of establishments that can be utilized when matching data of the statistical offices with data of the Federal Employment Agency. In some cases however, this identifier can be missing. Methodological prob-

[2] More details on data sets and the project as a whole can be found in Hethey & Spengler (2009) and Bender, Wagner, Zwick (2007).

lems also arise from the fact that under certain circumstances it may change over time. No such identifier is available when it comes to record linkage with data of the German Federal Bank. Hence, a big part of the data link-up has to be realized on the basis of company names, address information and variables which are available in all data sets like the legal form or the economic branch.

The initial situation in the project with respect to record linkage is favourable insofar as the overlapping variables are sufficiently broad. What's more, the variables are of high quality in all datasets. There is also a lot of literature from research on the topic. Knowledge about typical problems and necessary decisions is therefore available for many scenarios.[3] As for software, the decision has been made to use the Merge Toolbox developed by Rainer Schnell and others. A comprehensive overview over all the questions and issues that will arise in the context of data merging cannot be given here. Hence, just two points which both relate to the significance of name and address information will be made.

The name of a company is of great importance because one would not want to consider two records to constitute a match if there was not at least a medium level of concordance between the two name strings. As it is also true that string comparisons entail some challenges, it will be necessary to study this area and its problems in quite some detail. What kind of questions come up in this context can be highlighted by means of two examples:

- Company names often contain some general terms. At least in Germany, one often finds words like "Gesellschaft" (society), "Unternehmen" (enterprise, company) or abbreviations for the legal form like "GmbH & Co. KG". These words can be longer than the name of the company in a more narrow sense and less prone to different forms of spelling. Solutions for this matter have to be evaluated. The easiest approach is to delete such terms before string comparisons are carried out. However, this might not be the best strategy in all cases.
- Sooner or later, a decision with regard to the string comparison function has to be made. The number of potentially suitable functions is double digit. There is also a broad literature from several disciplines about what works best (cf. as a starting point Herzog et al. 2007). The reason for this is the fact that string comparisons are not only relevant in the context of record linkage but also in information retrieval, in all kinds of search processes respectively. Some string comparison functions which have proven to be powerful are the Jaro-Winkler-Similarity-Function (Jaro 1989, Winkler 2003), the Monge-Elkan-Distance (Monge & Elkan 1996), and the so called cosine similarity. What function will perform best under the specific circumstances in the KombiFiD project is not easy to tell.[4] To predict

[3] A useful overview can be found in the reports of the ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data (CEBEX-ISAD 2008a, 2008b) and in a recent book by Herzog et al. (2007). Detailed reports on what exactly has been done in certain empirical record linkage projects on company-level and to what extent certain things worked out, are – on the other hand – relatively difficult to get hold of.

[4] An empirical study done by Cohen et al. (2003) comes to the conclusion that the best results – at least in more complicated cases – are achieved with an approach that implements a combination of Jaro-Winkler and a scaled version of the Levenstein distance.

a best solution solely on the basis of theoretical considerations may even be impossible. A period of experimentation is likely to be necessary.

4. Empirical Results

In this section, preliminary results for the project related drop-out processes are presented (Some preliminary empirical results for the record linkage part and more details on the decisions made in this context will probably be available in September). As explained in section two, the current legal situation in Germany demands, that enterprises approve of a cross-institutional combination of their data. Details aside, the result is a two-level drop-out process: On the one hand, there are enterprises which cannot be asked any more (e.g. due to closure) or for which an approval for the link-up of certain data is practically impossible to obtain (e.g. due to restructuring that in the meantime has taken place).[5] On the other hand, there are cases in which enterprises – for whatever reasons – are not willing to support the project or cases in which there is no response. With regard to the question of how biased the data is, for which a record linkage can be carried out, it is not necessary to distinguish between these two processes. The following results are therefore always results of analyses in which enterprises with approval are compared with the original sample.

The main and overall result of the preliminary descriptive analyses is completely satisfactory. There are some differences and correlations – big enterprises, for instance, are more likely to give their consent than small ones. However, the overall distribution of assent over size classes, branches and regions is within reasonable parameters, and no area has been spotted yet, where approval collapses. That is to say, nothing has showed up so far, that couldn't be dealt with in a straightforward way. For illustrative purposes, a selection of two tables is shown below.

Table 1: Percentage of enterprises, which approved of the data matching by size class and economic branch (*first included year*; *preliminary results*)

Size class (employees)	Economic branch				Overall
	Manufacturing	Construction	Retail and wholesale	Services	
10/20 - 49	24,4	17,3	15,6	20,0	18,9
50 - 99	28,0	24,3	19,1	23,6	24,2
100 - 249	31,3	28,4	24,0	23,6	27,7
250 - 499	31,4	34,1	27,1	23,9	28,7
500 - 999	34,6	27,7	26,2	26,8	31,0
> 1000	35,3	19,2	36,5	27,3	33,4
Overall	28,7	20,8	18,1	21,3	22,4

[5] When it was already known in advance that a company is not active any more, the company wasn't included in the postal survey.

Table 2: Distribution of enterprises from manufacturing over economic branches: Comparison of the original sample and the group of approving enterprises (*column percentages*; *first included year*; *preliminary results*)

Economic branch (NACE 2003)	Original sample	Approving enterprises
C	1,4	1,5
DA	12,4	11,2
DB	5,3	4,6
DC	0,9	0,7
DD	2,6	2,5
DE	8,1	9,0
DF	0,3	0,4
DG	5,8	6,5
DH	5,3	5,5
DI	5,0	4,5
DJ	14,1	14,6
DK	15,6	15,7
DL	12,3	12,7
DM	5,8	5,8
DN	5,2	5,0
Overall	100	100

5. Outlook

The German KombiFiD project is still a work in progress. Important steps have already been taken. However, there is still some distance to go. A successful completion of the project would open up several possibilities: novel data would become available for scientific analyses without conducting additional surveys. Moreover, there would be perspectives for an improvement of the overall process in which official data on enterprises and their establishments is collected and processed. A further reduction of enterprises' reporting duties is one of the things that could be achieved.

References

Bender, S.; Wagner, J.; Zwick, M. (2007): *KombiFiD - Kombinierte Firmendaten für Deutschland*, Research Data Centres of the Federal Statistical Office and the statistical offices of the Länder, FDZ Working Paper No. 21 (also published as: University of Lüneburg, Working Paper in Economics Nr. 60 and Methodenreport 05/2007 of the Research Data Centre of the Federal Employment Service).

CENEX-ISAD (2008a): Report of WP1. *State of the art on statistical methodologies for integration 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.

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

Jaro, M. (1989): Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, in: *Journal of the American Statistical Association*, 84, 414-420.

Kohlmann, A. (2005): *The Research Data Centre of the Federal Employment Service in the Institute for Employment Research*, in: Schmollers Jahrbuch 125, 437-447.

Konold, M. (2007): New Possibilities for Economic Research through Integration of Establishment-level Panel Data of German Official Statistics, in: Schmollers Jahrbuch 127, 321-334.

Lipponer, A. (2003): *Deutsche Bundesbank's FDI micro database*, in Schmollers Jahrbuch 123, 593-600.

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.

Winkler, W. (1999): *The state of record linkage and current research problems*. U.S. Bureau of the Census. Statistical Research Division (PDF – R99-04).

Zühlke, S.; Zwick, M.; Scharnhorst, S.; Wende, T. (2004): *The research data centres of the Federal Statistical Office and the statistical offices of the Länder*, in: Schmollers Jahrbuch 124, 567-578.