Biewen, Elena; Gruhl, Anja; Gürke, Christopher; Hethey-Maier, Tanja; Weiß, Emanuel

Article

"Combined Firm Data for Germany" – Possibilities and Consequences of Merging Firm Data from Different Data Producers

Schmollers Jahrbuch – Journal of Applied Social Science Studies. Zeitschrift für Wirtschafts- und Sozialwissenschaften

Provided in Cooperation with:
Duncker & Humblot, Berlin

Suggested Citation: Biewen, Elena; Gruhl, Anja; Gürke, Christopher; Hethey-Maier, Tanja; Weiß, Emanuel (2012) : "Combined Firm Data for Germany" – Possibilities and Consequences of Merging Firm Data from Different Data Producers, Schmollers Jahrbuch – Journal of Applied Social Science Studies. Zeitschrift für Wirtschafts- und Sozialwissenschaften, ISSN 1865-5742, Duncker & Humblot, Berlin, Vol. 132, Iss. 3, pp. 361-377, https://doi.org/10.3790/schm.132.3.361

This Version is available at:
https://hdl.handle.net/10419/292373

Schmollers Jahrbuch 132 (2012), 361 – 377
Duncker & Humblot, Berlin

"Combined Firm Data for Germany" – Possibilities and Consequences of Merging Firm Data from Different Data Producers

By Elena Biewen, Anja Gruhl, Christopher Gürke, Tanja Hethey-Maier, and Emanuel Weiß*

Abstract

In the project "Combined firm data for Germany" (KombiFiD) firm data from different institutions were merged and made available for research for the first time. The institutions involved in the project faced considerable challenges both due to the narrow legal limits underlying such a merging procedure and as a result of the partial lack of a unique identifier. This paper provides an overview of the objectives associated with the project and its progress.

JEL Classification: C81

1. Introduction

In recent years the official statistics agencies have been able to expand their range of micro data in the field of enterprise and establishment data continuously in order to meet the scientific community's growing demand for micro data that are suitable for increasingly complex analyses.

The project "Combined firm data for Germany" (Kombinierte Firmendaten für Deutschland – KombiFiD), which has been running since January 2008, involves merging micro data at firm level beyond the boundaries of individual data producers. The institutions directly involved in the project are the Research Data Centre (FDZ) of the German Federal Employment Agency (Bundesagentur für Arbeit) at the Institute for Employment Research (IAB), Leuphana University of Lüneburg and the Research Data Centre of the Federal Statistical Office (Statistisches Bundesamt). In addition the Deutsche Bundesbank also puts parts of its pool of data into the project. During the course of the project a dataset was created which links firm data provided by various data

* The project "Combined firm data for Germany" (KombiFiD) is made possible by funding from the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung).

producers both with each other and over time. The result is a panel covering four years (from 2003 to 2006). In this way it was possible to demonstrate the fundamental legal and technical feasibility of the plan to generate a corresponding dataset comprising data from different institutions.

This paper first looks into the current situation regarding the range of firm data available to the scientific community and the possibilities for accessing them, and explains the objectives associated with the development of the KombiFiD dataset. As the merging of micro data collected by different data producers is subject to legal restrictions, in the third chapter the legal position is described and the resulting consequences are examined. This is followed by an examination of the methodological aspects of merging the micro data from the different institutions. A conclusion brings the paper to a close.

2. Currently Available Data

The research data centres provide micro data at both establishment and enterprise level for the scientific community via different routes of access (see Zühlke et al., 2003). There are, for instance, standardised scientific use files, whose micro data are de facto anonymous, which means that the potential costs of identifying an individual observation unit have to outweigh the benefits associated with this de-anonymisation (§ 16 para. 6 German Federal Statistics Law (Bundesstatistikgesetz – BstatG)). Furthermore, researchers also have the possibility to work with micro data on-site at guest researcher workplaces in the research data centres. In addition there is the possibility of remote data execution where the researcher does not come into direct contact with the data. Instead, he or she writes a code in a statistics program which is subsequently run over only formally anonymised data by a member of staff at one of the research data centres. No changes are made to these data apart from the removal of direct identifiers, such as the address of the company headquarters. The results obtained in this way are anonymised at the research data centres in such a way that it is no longer possible to re-identify individual observation units. The Deutsche Bundesbank also has micro data at enterprise level. These data can be used by researchers at the Research Centre of the Deutsche Bundesbank. In this case the data may only be used at guest researcher workplaces.

The firm data provided by the official statistics agencies already have a large information potential, partly due to the large sample sizes[1] and to firms' obligation to respond, which results in very low levels of non-response (see Brandt

[1] Many enterprise statistics are even exhaustive samples with cut-off thresholds, for example the investment surveys conducted among enterprises and establishments in manufacturing, mining and quarrying.

et al., 2008). Nonetheless it is necessary to continue to develop the micro data at establishment and enterprise level which is made available for research purposes, as the analysis possibilities in the scientific community are constantly expanding and the demand by policymakers for advice based on micro data is increasing continuously.

In the project "Official firm data for Germany" ("Amtliche Firmendaten für Deutschland" – AfiD) micro data from the Federal Statistical Office and the statistical offices of the German *Länder*, for example referring to enterprises in the services sector and establishments in manufacturing, were used to demonstrate that firm data gathered by the official statistics agencies can be prepared as a panel successfully in order to permit corresponding longitudinal analyses (see Malchin / Voshage, 2009). After establishment and enterprise data collected by one data producer had been linked over time, linking data from different data producers was the next logical step.

In the course of the KombiFiD project, firm-level micro data originating from the Federal Statistical Office and the statistical offices of the *Länder*, the Institute for Employment Research (IAB) and the Deutsche Bundesbank were merged.

The statistical offices make available various statistics from five fields (manufacturing, the construction industry, trade, the services sector, tax statistics). The Institute for Employment Research brings into the project the Establishment History Panel (Betriebs-Historik-Panel – BHP), which contains detailed information on employees. The data from the Bundesbank include financial statements of enterprises and data on foreign direct investment.

The statistics[2] from the different data producers which are integrated into the KombiFiD dataset are listed in Table 1.

As mentioned above, the linking of statistics from different data producers is associated among other things with the aim of being able to develop new analytical potential. It is conceivable, for example, that the KombiFiD dataset could be used to examine correlations between the development of direct investments and the employment structure, or between the cost structure of an enterprise and the development of employment (see L'Assainato, 2009). The current situation of firms having to respond to numerous surveys, sometimes containing the same questions, constitutes a burden both for the respondents and for the institutions collecting the data, although the obligation of industry to provide information for official statistics accounts for only a comparatively small part of the overall bureaucratic burdens faced by enterprises and establishments (see Vorgrimler et al., 2011). In addition, the dataset created in the KombiFiD project can be used to assess where there may be potential for rationalisation with regard to business surveys.

2 For a precise overview see www.kombifid.de.

Table 1

Statistics integrated into the KombiFiD dataset

Statistics from the RDCs of the statistical offices	Statistics from the IAB and the Deutsche Bundesbank
• Business Register System (URS) • Cost structure survey in manufacturing, mining and quarrying (KSE-VG) • Cost structure survey in the construction industry (KSE-Bau) • Structural survey in the services sector • Annual survey in wholesale and retail trade (and the maintenance and repair of motor vehicles and of personal and household goods) • Survey of investments in the main construction industry and in the finishing trade • Survey of investments in manufacturing, mining and quarrying • Monthly report incl. survey of new orders for local units in manufacturing, mining and quarrying • Turnover tax statistics • Annual report on enterprises in manufacturing, mining and quarrying • Structure of earnings survey in industry and the services sector	**Statistics from the IAB:** • Establishment History Panel **Statistics from the Deutsche Bundesbank:** • Foreign direct investment stock statistics • Financial statements of the Deutsche Bundesbank

Source: Own representation.

3. Current Legal Situation and the Resulting Consequences

According to the current legal situation[3] only data which are subject to a uniform legal basis may be merged. As the data from different data producers are collected on the basis of different legal bases, it is not possible to link data across institutional boundaries without obtaining express prior consent from the observation units concerned. The firms whose data were to be linked in the context of KombiFiD therefore had to be asked in writing for their consent for precisely this step. In order to obtain the most exact possible picture of all the enterprises contained in the initial statistics used for KombiFiD a sampling concept had to be designed which included enterprises from the economic sectors of manufacturing, services, trade and construction in the sample. Owing to

[3] See § 13a Federal Statistics Law (Bundesstatistikgesetz – BStatG) as of 7 September 2007

limited financial resources it was only possible to ask approx. 2% (approx. 55,000) of all the firms that were potentially eligible for inclusion in the planned KombiFiD dataset for their consent. Firms that had not responded to the letter were contacted in writing up to three times. This procedure resulted in a response rate of about 57%, with approximately 30% of all firms contacted giving their consent. Despite the rate of consent, which was clearly above the original expectations, the linked micro data show biases which can be attributed to selectivity. One source of selectivities can be found in a heterogeneous response behaviour, as can be seen from Tables 2 and 3.

Table 2

Response behaviour by economic sector and region as %

Headquarters of enterprise	Economic sector				Total
	Manufacturing	Construction	Trade	Services	
Western Germany	39.9	32.1	24.5	28.3	30.6
Eastern Germany	39.1	26.4	20.8	27.8	28.2
Total	39.9	30.7	23.8	28.2	30.7

Source: Own calculations.

Table 3

Response behaviour by economic sector and enterprise size class as %

Enterprise size class	Economic sector				Total
	Manufacturing	Construction	Trade	Services	
10/20 – 49	29.2	21.4	19.3	16.7	19.6
50 – 99	33.3	28.9	23.8	21.5	27.0
100 – 249	36.7	35.2	29.1	21.8	30.8
250 – 499	37.6	40.6	32.4	28.0	34.2
500 – 999	42.2	34.0	34.3	33.1	38.4
> = 1000	42.0	19.2	38.9	32.4	38.5
Total	34.2	25.4	22.1	18.7	30.7

Source: Own calculations.

In order to tap the additional analytical potential that could be provided by a dataset containing micro data from different data producers, there has to be the possibility in principle to be able to link firm data held by different institutions. This also applies for the utilisation of efficiency potentials associated

with reducing the mentioned multiple questions referring to the same variables: these too can probably only be fully exploited if the legal framework is amended accordingly.

For this reason, in the course of the project a legal opinion was sought with the aim of clarifying whether the described linkage of firm data can be implemented in the long term without having to obtain separate express consent from the enterprises concerned, and if so, under what circumstances. It must be emphasised that § 13a of the Federal Statistics Law (Bundesstatistikgesetz – BStatG) already constitutes the basis for a possible amendment to the law. In the sense of the above-mentioned aims of reducing respondent burden for firms in the long run and improving the use of and access to data for empirical research, both an extension of the scope of § 13a BStatG to cover the relevant datasets of the Federal Employment Agency and the Deutsche Bundesbank as well as the permanent admissibility of linking micro data across institutional boundaries would be desirable. According to the legal assessment, an extension of § 13a BStatG to include the permanent establishment of the possibility to link business micro data across institutional boundaries does not conflict in principle with the constitutional guidelines.

4. Linking the Data from the Federal Statistical Offices and the Statistical Offices of the *Länder*, the Institute for Employment Research (IAB) and the Deutsche Bundesbank

The KombiFiD project was conducted in two steps. In the first step the data from the Federal Statistical Office and the statistical offices of the *Länder* were combined with the Establishment History Panel of the IAB. The data were linked via the BA establishment numbers, which are contained in both the Business Register System (Unternehmensregister – URS) of the Federal Statistical Office and the Establishment History Panel (Betriebs-Historik-Panel – BHP) of the IAB. In the second step the data from the Bundesbank were merged. As the datasets of the Bundesbank do not have any numerical identifiers in common with the data of the other partners in the project, the data were linked in this case by comparing the names and addresses of the businesses. In sections 4.1 and 4.2 the procedure used to link the data from Federal Statistical Office and the statistical offices of the *Länder* with those from the IAB for KombiFiD Versions 1.0 and 2.0 is described.[4] Section 4.3 describes the subsequent linkage with the Bundesbank data.

[4] For Version 2.0 an attempt was made to increase the number of firms for which a link can be made between the data of the Federal Statistical Office and the statistical offices of the *Länder* and the IAB. For this reason the linkage procedure was modified.

4.1 Linking the Data for KombiFiD Version 1.0

The result of the steps described below is a key file which simplifies the matching of the datasets.

On the part of the Institute for Employment Research, the individual data of employees covered by social security, which are aggregated to establishment data in the Establishment History Panel (Betriebs-Historik-Panel – BHP), are made available for the KombiFiD dataset.[5] In addition to basic information about employee, age and wage structures, the dataset also contains variables concerning inflows and outflows of employees.

In contrast to the BHP, the KombiFiD dataset depicts the enterprise level. This makes two essential demands on the linkage key used for merging the IAB data with the datasets of the Federal Statistical Office and the statistical offices of the *Länder*. First, an identifier is required which makes it possible to match the units of the different datasets clearly with each other. Second, an aggregation key is needed that permits the aggregation of the establishment data from the BHP to the enterprise level. Both of these requirements are fulfilled by the Business Register System (URS). This contains both the BA establishment number, with the aid of which it is possible to match clearly the units of the BHP with those of the statistical offices, and an enterprise number, which permits the aggregation of the establishment units to the enterprise level.[6] The URS was thus used as the master file for the linkage procedure.

The URS extract used for the linkage contains the establishment and enterprise numbers of all firms that agreed to their data being merged. The data were delivered in individual cross-sections for the years 2003 to 2008.[7] Prior to the matching process the URS data were prepared for the requirements of the linkage. Units with no BA establishment number recorded in the URS were deleted, as a link with the BHP was not possible in these cases. Enterprises that could not be identified completely in the BHP, i.e. not with all the establishment numbers recorded in the URS, were also deleted before the linking procedure. This means that the linkage was limited to enterprises that were found in the BHP "in their entirety", in other words with all their establishment units. Establishments characterised by identical BA establishment numbers and enterprise numbers occurring more than once were also deleted prior to matching.

---

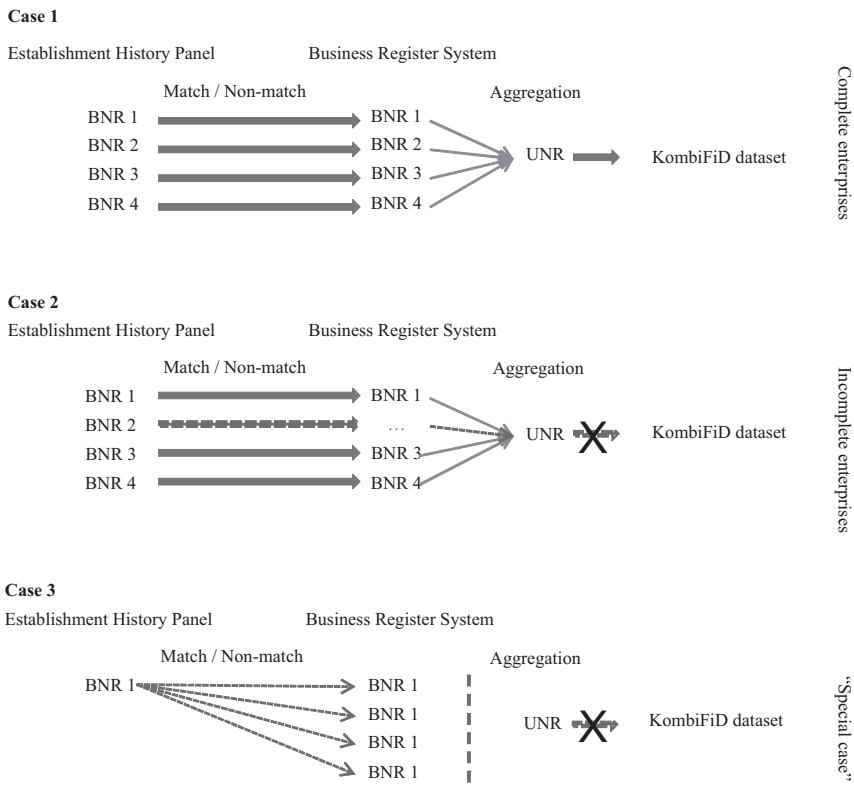
5 For more detailed information on the BHP, see Spengler (2008).

6 In the Business Register System each establishment number (Betriebsnummer – BNR) is assigned to an enterprise number (Unternehmensnummer – UNR). The data on all establishment numbers which are located under one specific enterprise number are combined and incorporated into the KombiFiD dataset in this form (e.g. total number of employees).

7 With the linkage to the BHP it was necessary to take a two-year time lag into account. The reason for this is that qualitatively sound data from administrative sources are available for the reporting year two years previously (Statistisches Bundesamt, 2009).

These establishments cannot be identified in the BHP as they are subsumed under one establishment number. In contrast to this, the units of such special cases are recorded separately in the URS. This leads to identical establishment numbers possibly occurring more than once in the URS.[8]

The key file resulting from the linkage procedure, which is used for linking the BHP with the data of the Federal Statistical Office and the statistical offices of the *Länder,* therefore only contains identifiers for firms with a full set of establishment numbers.



Source: Own representation.

Figure 1: Schematic representation of the process to link the URS and the BHP, KombiFiD Version 1.0

After the data preparation the BHP and the URS extract were matched using the BA establishment number as a unique numerical identifier. Then the estab-

[8] For detailed information see FDZ-Methodenreport 01/2010, http://fdz.iab.de/187/section.aspx/Publikation/k100311r01.

lishment data were aggregated to enterprise level on the basis of the establishment numbers and the associated enterprise numbers.

Figure 1 illustrates these clean-ups, the linkage and the aggregation in a highly simplified form.

Linking the URS and the BHP yielded very good results. The proportions of URS enterprises that could be identified completely in the BHP are clearly above 90% per cross-section examined, and are thus in a range which can be rated as very positive for the analysis possibilities using the KombiFiD dataset. Approximately 80% of the firms could be observed during the entire observation period. Merely three percent of the enterprise numbers identified appeared in only one cross-section. The vast majority of the enterprises from the first year of observation exist throughout the entire period that is of relevance here including all of their local units (establishments). In the last one-year cross-section 87.8% of the enterprises from the first year of observation were still depicted in the data in their entirety.

Table 4

Number of complete enterprises identified in the BHP and single-establishment enterprises

year, referring to URS extract	*Complete* enterprises identified in the BHP		Single-establishment enterprises	
	absolute	percentage of all *complete* URS enterprises	absolute	percentage of all URS enterprises
2003	13,296	96.2	11,994	73.9
2004	13,600	96.1	12,158	71.4
2005	13,722	95.7	12,256	71.7
2006	13,653	95.0	12,173	71.7

Source: KombiFiD data; own calculations.

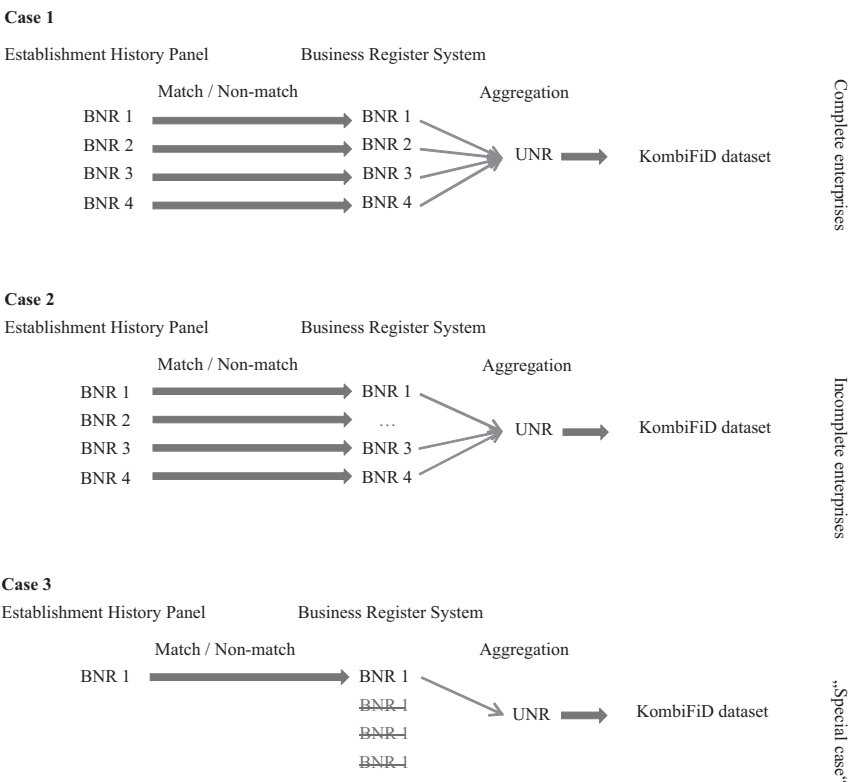
4.2 Linkage of KombiFiD Version 2.0

In contrast to the linkage procedure for KombiFiD Version 1.0, incomplete enterprises and the cited special cases were also taken into account when creating KombiFiD Version 2.0. However, it is still possible to identify the group of so-called complete enterprises. This is done via generated additional variables in the BHP spectrum of variables.

As was done for Version 1.0, units with no BA establishment number recorded in the URS (missings) were deleted, as in these cases a link with the BHP was not possible. In the case of enterprises for which it was not possible

to identify all the associated establishment numbers in the BHP, the establishment numbers that could not be matched were deleted after the linkage procedure. The linked establishment numbers of these enterprises were subsequently aggregated. This procedure differs from that used to create Version 1.0 of the KombiFiD dataset, where such incomplete enterprises were deleted entirely from the dataset. In cases where an establishment number was reported more than once in the URS, all entries of the establishment number apart from one were deleted. This remaining establishment number was linked with the corresponding unit in the Establishment History Panel. This, too, is a deviation from Version 1.0, in which these special cases were generally deleted and not taken into consideration for the linkage procedure.

After the data had been prepared, the BHP and the URS extract were matched using the BA establishment number as a unique numerical identifier. Then the establishment data were aggregated to enterprise level on the basis of the establishment numbers and the associated enterprise numbers.



Source: Own representation.

Figure 2: Schematic representation of the process to link the URS and the BHP

Figure 2 illustrates these clean-ups, the linkage and the aggregation in a highly simplified form.

For each one-year cross-section it was possible to identify between 85% and 88% of the enterprises in the URS entirely in the BHP. Between 75% and 79% of all enterprises are so-called single-establishment enterprises. Table 5 shows the results for the individual cross-sections of the KombiFiD data.

Table 5

Number of complete enterprises identified in the BHP and single-establishment enterprises

year	total number of enterprises	*Complete* enterprises identified in the BHP		Single-establishment enterprises	
		absolute	percentage of all URS enterprises	absolute	percentage of all URS enterprises
2003	15,812	13,722	86.8	12,256	77.5
2004	15,888	13,653	85.9	12,173	76.6
2005	15,924	13,580	85.3	12,077	75.8
2006	15,942	13,549	85.0	12,019	75.3

Source: KombiFiD data; own calculations.

4.3 Linkage with the Deutsche Bundesbank Data

After the data of the statistical offices had been linked with the data of the IAB as described above, in a further step two datasets of the Deutsche Bundesbank, the Microdatabase Direct Investment (MiDi) and the corporate balance sheets (USTAN), were added. The MiDi contains information about German direct investment abroad (outward FDI) and foreign direct investment in Germany (inward FDI) when certain reporting thresholds are exceeded (see Lipponer 2003, 2009). The USTAN comprises the financial statements of non-financial companies which the Bundesbank receives in the context of the refinancing business (see Stöss 2001). In contrast to the first linkage, in this case there was no common error-free key shared by the Deutsche Bundesbank, the statistical offices and the IAB. *Record linkage* techniques were therefore resorted to here, which permit a link even when keys contain errors[9]. Names and addresses in the datasets of the Deutsche Bundesbank and the IAB served as keys here. The matching process was implemented in technical terms using the record linkage software *MTB* (*Merge-Toolbox*) which was developed at the University of Duisburg.[10]

[9] An insight into the subject of record linkage is provided by Winkler 1995.

[10] A detailed description of the software can be found in Schnell et al. 2005.

An address file with the names and addresses of all enterprises from the MiDi and USTAN as of 2006 was made available by the Deutsche Bundesbank. For data protection reasons this file was padded with further enterprise addresses from other enterprise databases held by the Deutsche Bundesbank (e.g. DAFN E, Hoppenstedt) before it was transmitted to the IAB. The file thus contained a total of 76,051 entries. The IAB prepared their so-called establishment file as of 2006. This was an address file of all establishments in Germany that were active in 2006 and had at least one employee covered by social security or in marginal part-time employment. The file contained 2,734,332 entries. As was the case with the link between the data of the statistical offices and those of the IAB which was described above, here too there was the problem that the IAB dataset did not refer to the enterprise level but to the more finely structured establishment level. It was known from previous studies, however, that in the case of enterprises with employees covered by social security the address of the enterprise is generally also the address of an establishment. During the linkage procedure an attempt was therefore made to allocate an establishment address from the IAB establishment file and therefore also an establishment number to each of the enterprises from the Deutsche Bundesbank dataset. The URS extract, which was described earlier and contains all of the establishment numbers for each of the firms that gave their consent to the merging of data, was subsequently used to determine whether the establishment and thus the enterprise behind it is included in the KombiFiD sample.

Before the actual link was conducted, the datasets were standardised to a large extent in a *preprocessing* stage in order to minimise different spellings of the same names and addresses. The elements of this preprocessing work included converting all letters into capital letters, recoding umlauts (Ä = AE, …), deleting special characters (e.g.: > ! / ;), standardising address components (e.g. Straße, Weg), deleting spaces and correcting incorrect postcodes. In this way at the end of the preprocessing stage the names and addresses were available on both sides in the following variables: enterprise name, legal form, street, street number, town, postcode.

Before applying error-tolerant linking procedures the two datasets were subjected to a precise comparison. For this an observation from the Deutsche Bundesbank dataset had to correspond 100 percent to an observation from the IAB dataset in selected combinations of variables in order to be classified as a valid link. In order to reduce the computing time for this comparison, a so-called "blocking" procedure was used. Here not all of the observations of the two datasets are compared with each other but only those which have identical entries on the blocking variable. The two- or three-digit postcode was used as the blocking variable. Table 6 lists the three different combinations of variables that were used in the given order for the precise comparison, as well as the number of valid links that resulted from the procedure.

Table 6

Results of the precise comparison of names and addresses

Model	Blocking variables	Variables	Number of valid links
1	3-digit postcode	Enterprise name, legal form, town, street, street number	22,080
2	3-digit postcode	Enterprise name, legal form, 5-digit postcode, street, street number	631
3	2-digit postcode	Enterprise name, legal form, town, street, street number	20

After completing the first three precise comparisons it was thus possible to assign 22,731 enterprises (approx. 30 percent) in the Bundesbank dataset to an establishment number.

The precise comparison was subsequently replaced by the use of an error-tolerant similarity function. In this case the so-called bigrams[11] were used when comparing the enterprise names. It was decided not to use any further similarity functions with other link variables and blocking was performed restrictively at street number or street level. In this way it was possible to link correctly even enterprises in industrial estates and large office blocks where all the firms have the same postal address and only differ from one another in the enterprise name. Table 7 shows the results of the error-tolerant comparisons conducted in the given order.

Table 7

Results of the error-tolerant comparison of names

Model	Blocking variables	Variables	Similarity function	Number of valid links
4	5-digit postcode, street, street no.	Enterprise name	Bigrams	15,435
5	2-digit postcode, town, street, street no.	Enterprise name	Bigrams	114
6	5-digit postcode, street	Enterprise name	Bigrams	2,976

After completing the error-tolerant comparison a total of 41,256 enterprises (approx. 55 percent) in the Bundesbank dataset could therefore be assigned to

[11] For this the enterprise name is split into strings consisting of two characters (bigrams). The number of matching bigrams on both sides determines the similarity of the observations.

an establishment number. The search for these establishment numbers in the URS key revealed that 4170 of these enterprises are included in the KombiFiD sample.

The KombiFiD enterprises identified after comparing the addresses were then linked with the two Deutsche Bundesbank datasets, the Microdatabase Direct Investment (MiDi) and the corporate balance sheets (USTAN). Before the linkage procedure, duplicate records were first eliminated. Duplicate records are cases where a URS enterprise number had been assigned to more than one enterprise in the Bundesbank data. This occurred particularly frequently when in addition to a firm there was also a holding company with the same name. As holding companies have their own enterprise numbers in the Bundesbank data but frequently do not show any employees covered by social security of their own, they were linked with the firm of the same name under which employees covered by social security were registered in the IAB address file. After such cases had been checked manually, the number of enterprises included in the KombiFiD sample decreased from 4170 to 3788.

Table 8 shows the number of enterprises in the KombiFiD sample about which information is available in the Microdatabase Direct Investment (MiDi). Depending on the year the firm numbers vary between 702 (in 2003) and 780 (in 2005), with 681 enterprises appearing in all four years in the 2003 – 2006 period ("panel" line in the table). Consequently Table 8 shows numbers of MiDi firms that could be linked with the Establishment History Panel (Betriebs-Historik-Panel – BHP) of the IAB and with selected statistics of the Federal Statistical Office and the statistical offices of the *Länder*. Approximately 97% of the firms can be found in the BHP and about 76% in the turnover tax statistics. Furthermore, manufacturing is the best covered sector. Approximately 63% of the enterprises in the MiDi can be found in the monthly report, about 60% in the cost structure survey, in the annual report for multi-unit enterprises and in the survey of investments. The linkage with the statistics from other economic sectors (trade, services, construction) was far weaker, however.

Numbers of linked enterprises alone are not sufficient to make any statements regarding the quality of the dataset. For this reason in the next work step replication studies and further quality analyses are conducted to test whether analyses based on original data resemble those based on the KombiFiD sample.

Table 8

Linked MiDi enterprises

Year	MiDi	Link between MiDi and								
		BHP	MB VG	KSE VG	JB VG	IE VG	JE H	USS	SE D	VSE
2003	702	675	416	410	402	404	115	546	90	
2004	732	716	447	430	437	437	117	565	96	
2005	780	772	475	461	463	466	129	595	106	
2006	771	769	472	459	455	461	130	578	100	268
Panel	681	651	388	377	373	493	110	496	84	

BHP: Betriebs-Historik-Panel (Establishment History Panel); MB VG: Monatsbericht im Verarbeitenden Gewerbe (Monthly report in manufacturing, mining and quarrying); KSE VG: Kostenstrukturerhebung im Verarbeitenden Gewerbe (Cost structure survey in manufacturing, mining and quarrying); JB VG: Jahresbericht für Mehrbetriebsunternehmen im Verarbeitenden Gewerbe (Annual report on multi-unit enterprises in manufacturing, mining and quarrying); IE VG: Investitionserhebung im Verarbeitenden Gewerbe (Survey of investments in manufacturing, mining and quarrying); JE H: Jahreserhebung im Handel (Annual survey in wholesale and retail trade); USS: Umsatzsteuerstatistikpanel (Turnover tax statistics panel); SE D: Strukturerhebung im Dienstleistungsbereich (Structural survey in the services sector); VSE: Verdienststrukturerhebung im Produzierenden Gewerbe und im Dienstleistungsbereich (Structure of earnings survey in industry and the services sector) (cross-section 2006). Case numbers for other statistics are lower and are not reported here for reasons of space.

Source: Own calculations.

5. Conclusion of the "KombiFiD" Feasibility Study

The KombiFiD feasibility study laid the foundations for new considerations with regard to a possible reduction of respondent burden for businesses obliged to provide information and with regard to new analysis possibilities for the scientific community conducting empirical work. In future in particular the demand for an amendment to §13a of the Federal Statistics Law to create a permanent legal basis for the linking of business data across the boundaries of data producers should be in the focus of all those involved, the data producers, associations and interest groups, policymakers and the scientific community. A relevant amendment would fulfil two aims. Dropping the obligation to obtain consent from the observation units for the merging of their data, for instance, would be accompanied by the elimination of those potential selectivity effects associated with the heterogeneous response behaviour described above. Furthermore, writing to all the enterprises that are eligible for a merging of their micro data held by different data producers involves considerable resources. Therefore the mentioned amendment to the law would be accompanied by large

efficiency gains in any future projects aimed at creating datasets containing data from several different data producers.

Irrespective of the creation of a legal basis, the harmonisation of terms and methodologies between the national data producers is also an important milestone for creating a high-quality and comparable databasis of micro data from different data producers for the scientific community conducting empirical work.

Further quality analyses referring to the selectivity effects affecting the KombiFiD dataset are to follow, which will bring about learning effects for future non-compulsory surveys.

The present KombiFiD dataset, Version 2.0, is available for use by the scientific community until the end of 2021, legal restrictions having prevented a longer period of use. The relevant applications for data use and the rules underlying access to the data, as well as the meta data can be found under www.kombifid.de.

References

Brandt, M./*Oberschachtsiek,* D./*Pohl,* R. (2008): Neue Datenangebote in den Forschungsdatenzentren – Betriebs- und Unternehmensdaten im Längsschnitt, AStA Wirtschafts- und Sozialstaatliches Archiv 2, 193–207.

L'Assainato, S. (2009): KombiFiD – Kombinierte Firmendaten für Deutschland: Institutionenübergreifende Zusammenführung von Unternehmensdaten, DRV-Schriften, Band 55, 39–54.

Lipponer, A. (2003): Deutsche Bundesbank's FDI Micro Database, Schmollers Jahrbuch/Journal of Applied Social Science Studies 123, 593–600.

Lipponer, A. (2009): Microdatabase Direct Investment – MiDi a brief guide, Deutsche Bundesbank, http://www.bundesbank.de/vfz/vfz_forschungsdaten_einzeldaten.php.

Malchin, A./*Voshage,* R. (2009): Official Firm Data for Germany, Schmollers Jahrbuch/Journal of Applied Social Science Studies 129, 501–513.

Schnell, R./*Bachteler,* T./*Reiher,* J. (2005): MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung, ZA-Information 56, 93–103.

Spengler, A. (2008): The Establishment History Panel, Schmollers Jahrbuch/Journal of Applied Social Science Studies 128, 501–509.

Spengler, A. (2010): Verknüpfung und Abgleiche von Unternehmensregisterdaten des Statistischen Bundesamtes mit Betriebsdaten des Instituts für Arbeitsmarkt- und Berufsforschung, FDZ-Methodenreport 1, Nürnberg.

Stöss, E. (2001): Deutsche Bundesbank's Corporate Balance Sheet Statistics and Areas of Application, Schmollers Jahrbuch/Journal of Applied Social Science Studies 121, 131–137.

Vorgrimler, D./*Spengler,* F./*Schüßler,* S. (2011): Konzeption und erste Ergebnisse des Belastungsbarometers für Wirtschaftsstatistiken, Wirtschaft und Statistik 6, 528–535.

Winkler, W. E. (1995): Matching and Record Linkage, in: Cox, B. G. et al. (eds.), Business Survey Methods, New York, 355 – 384.

Zühlke, S. / *Zwick,* M. / *Scharnhorst,* S. / *Wende,* Th. (2003): Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, Wirtschaft und Statistik 10, 906 – 911.