Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ-Methodenreport

Methodological aspects of labour market data

04/2018
EN

Orbis-ADIAB: From record linkage key to research dataset

Combining commercial company data with administrative employer-employee data

Manfred Antoni,
Katharina Koller,
Marie-Christine Laible,
Florian Zimmermann

Bundesagentur für Arbeit

Orbis-ADIAB: From record linkage key to research dataset

Combining commercial company data with administrative employer-employee data

Documentation version: DOI: 10.5164/IAB.FDZM.1804.en.v1

Manfred Antoni (Institute for Employment Research)

Katharina Koller

Marie-Christine Laible (Institute for Employment Research)

Florian Zimmermann

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Contents

List of tables

Abstract
We describe the first steps of creating a research dataset from a company-establishment correspondence table. This correspondence table allows the combination of commercial company data from Bureau van Dijk (BvD) with administrative employment data of the Institute for Employment Research (IAB). To combine the research potentials of both data sources, a record linkage of independent companies given in BvD's database Orbis with their dependent subunits, i.e. establishments given in the Establishment History Panel (BHP), was performed. We describe the major challenges in creating the research dataset from the record linkage key. We then present representativity analyses that examine the selectivity of the resulting linked dataset. The new research dataset contains longitudinal information on companies, their dependent establishments and their employees. This dataset is currently in an internal test phase and cannot yet be accessed by the research community.

Zusammenfassung
Wir beschreiben den Prozess der Erstellung eines Forschungsdatensatzes aus einer Korrespondenz-Tabelle zwischen Unternehmen und ihren Betrieben. Diese Korrespondenz-Tabelle erlaubt eine Verknüpfung kommerzieller Unternehmensdaten von Bureau van Dijk (BvD) mit administrativen Arbeitsmarktdaten des Instituts für Arbeitsmarkt- und Berufsforschung (IAB). Um die Forschungspotentiale beider Datenquellen zu verbinden, wurden unabhängige Einheiten, d.h. Unternehmen der Datenbasis Orbis von BvD, mit den zugehörigen abhängigen Einheiten verknüpft, d.h. mit Betrieben des Betriebs-Historik-Panels (BHP). Die wesentlichen Herausforderungen der Erstellung des Forschungsdatensatzes aus dem Record-Linkage-Schlüssel werden beschrieben. Des Weiteren präsentieren wir Repräsentativitätsanalysen, die die Selektivität des resultierenden verknüpften Datensatzes untersuchen. Dieser neue Forschungsdatensatz enthält Längsschnittinformationen über die Unternehmen, den dazugehörigen Betrieben und allen Beschäftigen in diesen Betrieben. Momentan befinden sich die Daten in einer internen Testphase und stehen der Wissenschaftsgemeinschaft noch nicht zur Verfügung.

Keywords: linked commercial-administrative data, linked employer-employee data, record linkage, company data

Disclaimer:

Orbis-ADIAB is in an ongoing test phase cannot be accessed by the research community.

The authors would like to thank Dana Müller, Alexandra Schmucker, Heiko Stüber, Simon Trenkle, Benjamin Wirth, the participants of the Fifth International Conference on Establishment Surveys (ICES-V) and of the Administrative Data Research Network Conference (ADRN) 2017.

1 Introduction

The Research Data Centre (FDZ) of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB) acquired a data extract from the commercial data provider Bureau van Dijk (BvD). BvD's company-level financial data were linked to administrative data of the BA (Schild, 2016) to obtain a key between company identifiers in the BvD data and establishment identifiers of the IAB. The key is then used to link BvD's database Orbis to the Establishment History Panel (BHP). The report at hand describes the challenges and steps taken to generate a valuable research dataset from this key. The resulting research dataset is henceforth called Orbis-ADIAB.

BvD's main business objective is to provide business intelligence for private companies. Therefore, BvD follows commercial interests, as opposed to research goals, which is reflected in Orbis' data structure. While BvD's database Orbis contains financial information on companies worldwide, the FDZ's extract of Orbis is restricted to companies in Germany. The data used for record linkage was drawn on January 30th 2014[1] and covers companies' background information, yearly financial information and quarterly financial information for stock traded companies.

While the information made available by BvD is valuable in itself, enhancing the dataset by performing record linkage leads to higher information density. Thus far, no systematic large-scale linkage between BvD data and administrative data has been realized for Germany. Such a linkage increases research opportunities and presents a unique value added: Establishment information inherent to IAB's data products, e.g. the BHP, can be combined with information on the overlying company for the first time. Up to this point it was not possible to aggregate establishment-level data (dependent subunits) to company-level data (independent unit).[2] Additionally, the combined financial-administrative data create the possibility to analyze data on three levels: the company-level, the establishment-level and the employee-level. Overall, a combination of the financial and administrative data creates a new dataset that is specifically suited for scientific research, as opposed to commercial interests.

2 The datasets linked by the key

The key generated by record linkage brings company information from BvD together with establishment information from the BHP. The main features of the two datasets are described below.

2.1 Bureau van Dijk

Bureau van Dijk (BvD) is a commercial provider of company data and business intelligence. Founded in 1971, BvD's core competencies include the collection and compilation of publicly available company data from different sources in order to supply comprehensive information

[1] Two updates on January 30th of 2015 and 2016 are also available to the FDZ and the research data have been updated to include financial information from these updates.

[2] Biewen et al. (2012) and Schäffler (2014) describe two projects which also linked external company data to administrative microdata of the IAB. However, these linked data are either no longer available to the scientific community or never have been available to researchers outside the IAB.

for private companies. BvD's databases have mainly been used for analyzing business intelligence and are well renowned with companies researching competitors, suppliers or business partners. More recently, BvD's databases have additionally become known to academic researchers.

Currently BvD offers over 20 different databases containing information on a worldwide basis, as well as country-specific data. Of these databases, Orbis contains financial information for companies worldwide. For the purpose of the record linkage, only information on German companies was used and this information is mainly sourced by BvD from the Creditreform Rating AG. A company in this context is defined as an independent unit with a specific legal form which incorporates one or more establishments.

The data extract used for the record linkage contained a total of 1,948,778 companies on the reference data of January 30th 2014 of which 535,129 were matched with data from the IAB. The data are split into three parts: One dataset contains background information on the reference date. The second dataset includes yearly financial information from 2006-2013 for all German companies for which this information is available, while the third dataset contains quarterly financial information for stock traded companies only. Among the variables are assets, debts and liabilities, capital, revenue, sales, profit, costs and EBITDA[3], as well as industry and location information. The timeframe was chosen according to the availability of the data, as BvD's financial information is most reliable since 2006, due to changes in the German financial reporting system. The year 2006 thus is the earliest year for which BvD's Orbis[4] data can be merged with the administrative Establishment History Panel from the IAB.

2.2 Administrative data — Establishment History Panel (BHP)

The Establishment History Panel (BHP) comprises all German establishments with at least one employee subject to social security contributions on the reference date of June 30th of each year. Since 1999, establishments with at least one marginal part-time employee are also included in the BHP. According to the BA's definition, an establishment is one or more branch office(s) of a company which is assigned a specific establishment number ("Betriebsnummer"). At least one establishment number is assigned for each municipality the company holds a branch in. Additional establishment numbers per company and municipality are issued if different branches within the same municipality follow different economic purposes and therefore belong to different sectors.

The BHP is made up of yearly cross-sectional datasets starting in 1975 in Western Germany and 1992 in Eastern Germany. In 2013, the cross-section contained over 2.9 million establishments. The individual cross-sections can be combined into a panel dataset using a unique and time-consistent establishment number. For Orbis-ADIAB, we merge the BHP of the years 2006 to 2013 with BvD's Orbis data.

[3] A company's earnings before interest, taxes, depreciation, and amortization.
[4] In the remainder of the text we simply refer to our data extract as Orbis, although the extract only contains the subset of companies in Germany as of January 30, 2014.

The BHP contains information on the industry sector and the location of an establishment, as well as the number of employees liable to social security and marginal employees. These employee numbers are further broken down by gender, age, occupational status, qualification and nationality. Aggregate information on wages is also available. For a detailed description of the BHP see Schmucker et al. (2016).

3 The record linkage key

Record linkage is the process of identifying records on a given observational unit in two or more different data files (Herzog et al., 2007). In this case, the goal is to match establishments as dependent subunits with companies as their superordinate units, thereby enabling research on two different levels. The company is the more aggregate unit and the record linkage process aims to assign one or more establishments to one overlying company in a one-to-many (1:n) relationship.

The FDZ's extract of Orbis contains non-unique linkage identifiers on the overlying company such as their name, legal form or address. Similar linkage identifiers on each establishment within the BHP were extracted from the Data Warehouse of the BA.

The challenge of this record linkage lies in identifying all matches between establishments which truly belong to same company. In order to meet this challenge, the linkage process relied on several identifiers, of which the name and the legal form were the most important (Schild, 2016). After extensive preprocessing of the identifiers, i.e. cleaning and standardizing the raw data to be more useful during the comparison and classification steps of the record linkage, different matching procedures were applied and 17 steps were taken in order to obtain the best possible result.[5]

For the years 2006 to 2013, at least one establishment could be assigned to 975,880 of the 1,948,788 companies (as of January 30th 2014) in FDZ's extract of Orbis. This translates into a matching success rate of 50% (Schild, 2016). Over 87% of all matches were found through exact matching — of the name in its long form or abbreviation in addition to the identifiers legal form, zip code, one-digit industry sector or place — indicating a high matching quality. Several plausibility checks further increase the confidence in the quality of the match (Schild, 2016).

Table 1 shows the matching success rate for all companies for the year 2006 to 2013 by different company size groups as defined by the variable measuring the number of employees (empl) in Orbis. Column (1) shows the share of cases where no establishment was matched to a company, and column (2) shows the share of successful matches. While, in absolute terms, more companies were assigned at least one establishment in the smaller size catego-ries, the matching success rate increases with company size, as comparatively fewer larger companies exist in Germany.

[5] For more details on the linkage identifiers, the preprocessing and the matching procedures, as well as on the overall record linkage process see Schild (2016). The record linkage was performed by the German Record Linkage Center (GRLC, see Antoni and Schnell (2017) or record-linkage.de), using methods developed by the GRLC and refined by Schäffler (2014).

Table 1: Matching success rate on the company-level by company size according to Orbis

Company size (in employees)	(1) No establishment assigned to company (in %)	(2) At least one establishment assigned to company (in %)	(3) Total
Missing	68.04	31.96	215,432
1-5	56.34	43.66	1,232,275
6-10	31.00	69.00	187,039
11-25	25.81	74.19	164,030
26-50	22.47	77.53	71,325
51-100	20.78	79.22	38,346
101-250	18.65	81.35	24,395
251+	20.34	79.66	15,936
Total	972,898	975,880	1,948,778
Total (in %)	49.92	50.08	100

Notes: When the company size (empl) was missing, information from older years was used if available.
Source: Own calculations based on the record linkage key and Orbis.

For the distribution of establishment-company matches see Table 2, which shows that for most companies with assigned establishments exactly one establishment match was found (84.3% of the 975,880 companies that are assigned at least one establishment). One has to be cautious however, as this does not imply that the particular company consists of only one establishment. A value of 1 in the ”number of assigned establishments” merely indicates that only one establishment could be assigned to a particular company through record linkage. We have no way of validating whether all German establishments of a given company were found.[7]

Table 2: Matches by number of assigned establishments for all companies

Number of assigned establishments	Number of companies	Companies (in %)	Share among linked companies (in %)
0	972,898	49.92	
1	822,871	42.22	84.32
2	107,889	5.54	11.06
3	22,921	1.18	2.35
4	8,065	0.41	0.83
5	3,963	0.20	0.41
6-10	6,191	0.32	0.63
11-100	3,720	0.19	0.38
101-1000	247	0.01	0.03
1001+	13	0.00	0.00
Total	1,948,778	100.00	100.00

Source: Own calculations based on the record linkage key.

[7] Orbis does not contain the number of establishments in Germany, so there is no benchmark against which we could validate the linked number of establishments per company. The only database that would allow such a validation for Germany is the Company Register (see https://www.unternehmensregister.de). For legal reasons, no extract of that register is available to us.

The final key resulting from the record linkage contains 1,386,439 establishments linked to a company in Orbis. Table 3 documents the activity status, i.e. whether a company is still active as opposed to inactive (insolvent or dissolved), and linkage status of the companies in our Orbis extract. Of 321,579 inactive companies, 40% were linked. For active companies the linkage was more successful, as 52% of these companies could be linked. This result is not surprising as the activity status indicates whether a company was solvent and still existed at the reference date January 30th, 2014.

Table 3: Number of active and linked companies

	Not linked (in %)	Linked (in %)	Total
Inactive	59.78%	40.22%	321,579
Active	47.98%	52.02%	1,627,199
Total	49.92%	50.08%	1,948,778

Source: Own calculations based on the record linkage key and Orbis.

4 The challenges of creating a dataset from the key

The successful record linkage of establishments to companies was the first step in creating a research dataset. Once the key, i.e. the company-establishment correspondence table, had been derived, a number of further steps had to be taken in order to obtain a valid dataset. This process included several challenges for which innovative solutions had to be found.

4.1 Dealing with multiple assignments

The first challenge provided by the key are m:n matches. Per definition, several establishments can be linked to one company (1:n), but an establishment can only be linked to one company (1:1). The key however contains matches where one establishment is matched to more than one company (m:1). This result is problematic, as it represents an impossible assignment according to the defined company-establishment relationship. Several steps (listed as MA1 to MA7) were taken to solve this problem, which are described in Table 4.

The first two steps of this process have the largest impact. In step MA1, we drop all matches for which we do not find corresponding financial information in the Orbis database.[8] This can be the case for companies who do not have to file an annual report or for whom information was not available to Orbis. The next step, MA2, concerns a corresponding procedure at the establishment level and all matches are dropped for which we do not find establishment-level information.

[8] We do this, as one prerequisite for a record pair's inclusion in the research dataset is the existence of information in both datasets.

Table 4: Steps taken to resolve multiple assignments

Step	Description
MA1	Merge with financial data from BvD, drop if there is no match in the financial data for at least one year from 2006-2013. The financial information is taken from the data extract of 2014 only and has to refer to the consolidation degree “unconsolidated” and the reporting standard “local GAAP”.
MA2	Merge the key with the BHP for the years 2006-2013, drop if no match is found. Only establishment-company links available in both datasets are retained.
MA3	Matches with an active company are favored over matches with a company with solvency problems as defined by BvD in the year 2014. Matches with an active company or a company with solvency problems are favored over matches with an inactive company.
MA4	Regional allocation is based on municipality names and zip codes in several iterative sub-steps. In the first sub-step, municipality names are compared. If at least one of the municipality names is missing, 5-digit zip codes are regarded. If municipality names or 5-digit zip codes correspond for more than one match, all matches with corresponding municipality names or corresponding 5-digit zip codes are retained. Likewise, if the municipality names or 5-digit zip codes do not correspond for any match, all matches are retained. In the second sub-step, the retained matches are then compared based on additional information pertaining to the zip code. Zip codes are considered to correspond to each other when the following conditions are met: 1) Zip codes are regarded with decreasing precision in the number of digits. 2) As zip codes may change over time, street name, house number and municipality name are compared. 3) Municipality names correspond and at least one zip code is missing. If these conditions are met for all matches, they are retained. If the conditions are not met for any of the matches, all matches are retained.
MA5	Matches with corresponding 1- to 5-digit industry code (establishment) or corresponding address (headquarters) are favored over non-corresponding industry code and non-corresponding address. Matches with a missing industry code do not get dropped.
MA6	If a company is only assigned one establishment, this link is favored over a link where more than one establishment is assigned to one company.
MA7	The remaining multiple assignments are dropped.

Notes: Local GAAP refers to the generally accepted accounting principles of Germany. Solvency problems refer to a variable available in Orbis. This variable contains the following solvency problems: The company is technically still active but in default of payment, dormant or in insolvency proceedings. The company is no longer active, i.e. dissolved by liquidation, merger or acquisition. Establishments with missing values in the variables used for comparison are not dropped in the relevant steps.

A considerable amount of matches is lost upon merging the key with the BHP (MA2). Several reasons were identified to explain this loss of potential establishment-company links. It is, for instance, possible that the registry for establishments is not updated in a timely fashion, such that establishments which are still marked as active in the Data Warehouse of the BA do not actually exist anymore. Therefore they were matched through record linkage processes using information on names and addresses, however, as the establishment itself no longer exists, no corresponding observation is found in the BHP. The existence of an establishment in the context of the BHP is determined by whether it has at least one person in dependent employment at June 30th of a given year.

Several other reasons are conceivable as to why no corresponding observations can be found in the address data and the BHP. First, data maintenance concerning notifications to the establishment registry may not be up to date. Second, some establishments procure establishment numbers which then remain unused. This is the case, for example, when an establishment number is requested, but the establishment does not have any employees subject to social security contributions for which notifications need to be made. Third, when the establishment number is allocated, the establishment itself may still not show up in the BHP due to not having employees subject to social security on the reference date (30th of June) during the first year. Fourth, establishments belonging to a larger entity may make joint notifications with other establishments of the same company. In this case several establishment numbers exist, but their employees are all registered under only one establishment number in the BHP, namely that of the notifying establishment. These notifications might also have been submitted by a tax accountant. The remaining steps have a lower impact on the resulting research dataset and are described in Table 4.

To illustrate the effectiveness of the steps described in Table 4, Table 5 shows the number of matches and the impact of each step. By impact we mean the number of changes a certain step would generate if the step were performed as the first step. This measure demonstrates the importance of any step on its own, irrespective of the order of the steps. The columns for (1) show the number of matches per step and the impact of the steps in turning multiple assignments from the line before into single assignments. The columns for (2) and (3) illustrate the results for the same steps for resolving double or triple assignments, i.e. when one establishment was linked to two or three companies. Table 5 shows that for the two subsets (2) and (3) the number of matches gradually decreases, which is a positive result as it shows that some multiple assignments have been resolved by the step at hand. The number of matches for the subset (1) decreases until MA4, then it starts to increase. This is due to the number of double- and triple assignments turning into single assignments.

In each step, double (triple) assignments are resolved and are turned into single (double) assignments. For example, looking at column (1), the overall resulting number of matches of N=861,786 is a combination of two simultaneous processes incurred during step MA1: First, the number of single assignments is reduced by dropping companies without financial data. Second, the number of single assignments is increased by resolving double or triple assignments. The impact for this step is a reduction in the number of matches by 503,537. Disregarding these matches in the final research dataset is necessary, as they are either false

matches or not meaningful for research purposes as they have no financial data content on the company side (MA1) or no establishment characteristics on the establishment side (MA2).

When looking at the columns for “matches”, we see that in step MA1 most of the double or triple assignments are turned into single assignments or dropped out of the dataset because they have no financial information. Further, the reduction of single assignments is highest in step MA2 because assignments are dropped due to missing establishment information.

Table 5: Resolving multiple company-establishment assignments: Matches and impacts

Steps	Matches and Impact					
	(1) Single assignment		(2) Double assignment		(3) Triple assignment	
	Matches	Impact	Matches	Impact	Matches	Impact
Orbis	1,948,778	0				
Linkage	1,365,323	0	41,754	0	717	0
MA1	861,786	-503,537	7,532	-34,222	51	-666
MA2	650,325	-401,025	6,838	-3,646	48	-27
MA3	650,878	8,456	5,740	-16,616	36	-444
MA4	652,195	10,119	3,116	-19,924	21	-471
MA5	652,784	8,520	1,944	-16,738	12	-453
MA6	653,197	4,755	1,124	-9,354	3	-234
MA7	653,197	NA	NA	NA	NA	NA

Notes: The number of single assignments increases from step MA3 onwards as double or triple assignments turn into single assignments.
Source: Own calculations based on the record linkage key.

The last row reveals the final number of 653,197 unique establishment-company matches in the research dataset. The last row, marked by “NA”, reveals that all problems were solved in steps MA1 - MA6. This result is due to the fact that the remaining multiple assignments are dropped in MA7. Overall, we conclude that the steps devised to solve the problem of multiple assignments are successful.

4.2 Assigning headquarters

For companies with more than one assigned establishment, a headquarters was chosen from the available establishments. Several iterative steps were taken in order to assign the headquarters, as described in Table 6. These steps mainly pertained to using information concerning addresses and industry codes, as well as company sizes. Note that we merely assign headquarters according to the information available to us. This does not mean that we know for certain if the establishment actually is a headquarters, it only implies that based on the information available to us, this establishment is the most likely choice to be a headquarters. As with previous decisions, we lack further data to validate whether or to which degree our approximation is accurate.

Table 6: Steps for assigning headquarters

Steps	Descriptions
H1	Establishment and company have corresponding municipalities. If the municipality is missing, corresponding zip codes are compared.
H2	Establishment and company have corresponding street names.
H3	Establishment and company have corresponding house numbers.
H4	Establishment which is closer to the company according to geocoding and routing is favored.
H5	Establishment which has been founded earlier according to the BHP is favored.
H6	Establishment with the highest administration portion is favored.
H7	Establishment with an industry code which is the most common among headquarters according to the IAB Establishment Panel Survey (Fischer et al., 2009) is favored.
Only for companies without an establishment with corresponding municipality or corresponding zip *code if municipality is missing:*	
H8	Establishment with corresponding 5-digit zip code is favored.
H9	Establishment with respectively 3-digit zip code is favored.
H5-7	The steps for H5-H7 above are repeated.

Notes: In the first three steps (H1-H3), establishments with missing information are not considered as headquarters if another establishment with matching information is linked to the company. From step H4 onwards establishments with missing values are considered as candidates for headquarters. The administration proportion is defined as the number of skilled commercial and administrative occupations and managers divided by the number of all employees.

Table 7 shows the effectiveness of the steps just described. For each step, the number of successful headquarters assignments is given. The first row (A) indicates the number of single-site companies (N= 494,771) for which only the first step had to be taken.

For multi-site companies (B, N=40,358), each step H1-H7 is taken consecutively for establishments for which a locality was assigned (B1). For companies without an establishment with corresponding municipality or, if the municipality is missing, corresponding zip code (B2), steps H8-H9 were taken and then steps H5-H7 were repeated for this subset of establishments.

In each step, multiple possible headquarters are resolved and turned into unambiguously assigned headquarters. For example, looking at row H2, the number 3,983 shows how many multiple possible headquarters have been resolved by this step. These have either been resolved into companies with an ambiguously or unambiguously assigned headquarters. The rows “previous steps unsuccessful” show that for 101 multi-site companies with an assigned

locality (within B1) and 110 multi-site companies without an assigned locality (within B2), respectively, not a single multiple headquarters could be resolved. In the end, the steps were successful for 445,537 single-site companies and 40,157 multi-site companies.

Table 7: Assignment steps of headquarters

Steps:	N	%	% Cum.
A Single-site companies	494,771		
H1	445,537	90.05	90.05
Locality not assigned	49,234	9.95	100
Total	494,771	100	
B Multi-site companies	40,358		
B1 Locality assignment			
H1	22,735	62.19	62.19
H2	3,983	10.90	73.09
H3	489	1.34	74.43
H4	354	0.97	75.40
H5	4,595	12.57	87.97
H6	4,222	11.55	99.52
H7	76	0.21	99.72
Previous steps unsuccessful	101	0.28	100
Total	36,555	100	
B2 No locality assignment			
H8	1,141	30.00	30.00
H9	260	6.84	36.84
H5	1,341	35.26	72.10
H6	887	23.32	95.42
H7	64	1.68	97.11
Previous steps unsuccessful	110	2.89	100
Total	3,803	100	
Total companies	535,129		

Source: Own calculations based on the record linkage key, the BHP and the IAB Establishment Panel Survey.

The types of headquarters assigned are presented in Table 8 for single-site companies and Table 9 for multi-site companies. The first row of Table 8 indicates that for 90% of the single-site companies, a corresponding municipality or 5-digit zip code (if the municipality was missing) was available. The remaining 10% neither had a corresponding municipality nor a matching 5-digit zip code. The headquarters of these establishments have most likely not been found by the record linkage.

Table 8: Final assignment of headquarters (single-site companies)

Type of headquarters	N	%
Locality corresponds to individual enterprise	445,537	90.05
Locality does not correspond to individual enterprise	49,234	9.95
Total	494,771	100

Source: Own calculations based on the record linkage key and the BHP.

Table 9 reveals that for most of the remaining 158,426 establishments belonging to 40,358 multi-site companies' headquarters were assignable. Column (1) shows the headquarters assignment of companies at the establishment-level, while column (2) shows the same for the company-level. An unambiguous headquarters can be assigned to 40,102 companies which were linked with overall 157,558 establishments. For 45 companies with 420 linked establishments, the headquarters cannot be assigned unambiguously. Thus, for most companies, we can assign a headquarters according to our heuristic. The remaining establishments had identical outcomes for each step, therefore we cannot make an unambiguous assignment of the headquarters. These establishments, with different establishment identifiers and belonging to the same company, could be located in the same building and make joint notifications (see section 4.1). Only 448 establishments linked to 211 companies were completely unaffected by our steps and an assignment was thus impossible.

Table 9: Final assignment of headquarters (multi-site companies)

	(1) Establishment-level		(2) Company-level	
Type of headquarters	N	%	N	%
Establishments linked to a multi-site company with unambiguous headquarters	157,558	99.45	40,102	99.37
Establishments linked to companies with ambiguous headquarters	420	0.27	45	0.11
Assignment impossible	448	0.28	211	0.52
Total	158,426	100	40,358	100

Source: Own calculations based on the record linkage key and the BHP.

4.3 Dealing with BvD-ID changes

Potential changes in company identifiers in Orbis (BvD-ID) were tracked through a key provided by BvD which described ID changes from one year to the next. One ID was retained per company in order to provide a consistent company identifier for all years. Acquisitions, which have been identified by clerical review (change of company name), have not had their ID changed.

4.4 Dealing with updates

Unfortunately, simply updating the data with each new data extract is not possible. On the one hand, BvD overwrites data during their weekly updates and on the other hand, the data provider Creditreform changed policies, such that only reports filed in the e-Bundesanzeiger can be provided by BvD for the data extracts from 2015 and 2016. We have therefore opted to use the information in the data extract of 2014 as a basis and to update missing information through

the extracts in 2015 and 2016. Note that this may mean that information per year per observation may not be sourced from the same data extracts.

Two further caveats concerning updates relate to the financial reporting framework and the degree of consolidation. While most financial information is based on the German generally accepted accounting principles (GAAP), some companies may report according to the international financial reporting standards (IFRS) as of their adoption by the European Union in 2002. Due to data updates, changes in the accounting practice could occur.[9]

Similarly, the degree of consolidation of the financial statement may change between the updates.[10] A consolidated statement refers to a statement of a company that integrates the statements of the mother company's subsidiaries. The method of integration can vary according to the importance of the interest owned by the parent in its subsidiaries. Unconsolidated statements are those which only pertain to the individual company, not the conglomerate.[11]

5 Structure of Orbis-ADIAB

Table 10 describes the structure of the newly generated dataset with all three levels of observations. The research data created pertain to the years 2006 to 2013, and Table 10 relates company-level (columns 1-3), establishment-level (columns 4-5), and individual-level (columns 6-7) information for each year.

Concerning the company-level, column (1) shows the number of companies with at least one corresponding establishment in the BHP per year. Under the condition that a match exists, column (2) shows the number of companies per year for which financial data is available in Orbis. The number of companies in column (2) decreases in the years 2012 and 2013 due to a lag in the reporting of financial reports and a lag in the reports being transmitted into the Orbis database. For comparison, column (3) reports the total number of companies with financial data available in our Orbis extract in the relevant year.

Column (4) indicates the number of establishments matched to one company. Note that the number of establishments is higher than their corresponding number of companies in (1) due to the definition of the linkage, i.e. more than one establishment can be matched to one firm. The fifth column supplies the number of establishments recorded in the BHP per year. A comparison of (4) and (5) gives a benchmark of the success of the linkage for each year.

[9] We cannot reproduce these changes.

[10] Note that the accounting framework and the consolidation code may also vary within a year. Specifically, the yearly financial data are based on the German GAAP standard. The data are unconsolidated. However, the quarterly financial data are only available for stock traded companies. Financial data for stock traded companies conforms to IFRS. The data are unconsolidated for these observations.

[11] Note that we can only indicate changes relating to the accounting framework or the consolidation level changes when this information is available in Orbis from one year to the next. However, not all changes to the database Orbis are tracked, especially not in between the reference dates of the extracts, therefore we cannot guarantee completeness of the information. When information was available, dummies were generated indicating these changes.

Column (6) indicates the number of workers employed in all establishment per year matched with a company and as a reference point, column (7) reveals the number of employees in all establishments available in the BHP.

Overall, Table 10 shows that the company-establishment linkage was rather successful. Roughly one sixth of all establishments are covered per year, while over a third of all employees are covered, giving further evidence for the successful linkage of larger firms.

Table 10: Structure of the datasets with three levels of observations

Year	(1) Matched companies with at least one establishment (BHP)	(2) Matched companies with financial data (Orbis)	(3) Companies with financial data (Orbis)	(4) Matched establishments (BHP)	(5) Establishments in total (BHP)	(6) Employees in matched establishments (BHP)	(7) Employees in total (BHP)
2006	367,260	381,075	1,027,181	432,674	2,729,252	11,791,445	32,056,316
2007	386,216	402,068	1,077,564	454,874	2,780,754	12,403,443	32,826,631
2008	405,525	443,559	1,119,546	475,550	2,804,470	12,967,072	33,388,527
2009	425,910	477,707	1,139,398	498,185	2,849,632	13,058,721	33,346,381
2010	448,607	496,717	1,130,194	523,239	2,890,234	13,570,542	33,841,174
2011	465,276	500,638	1,162,668	542,771	2,927,181	14,207,459	34,610,634
2012	464,992	226,443	595,860	543,863	2,950,533	14,466,127	35,048,465
2013	455,137	3,008	223,614	534,077	2,968,382	14,518,635	35,477,827
Total	535,129	535,129	1,948,778	653,197	4,746,794	/	/

Notes: The number of companies in (1) refers to all companies which have at least one valid establishment link in the relevant year. The number of companies in (2) refers to all companies which have valid financial data in the relevant year. The numbers in (1) and (2) differ, as the company may have existed (and reported financial information) before the foundation of a linked establishment. Likewise, a company-establishment link may exist in the relevant year, but the company may not have filed an annual report. The number of employees in (6) and (7) has been calculated as the sum of total employees across all establishments. Thus a total number of employees cannot be calculated.
Source: Own calculations based on the record linkage key, the BHP and Orbis.

Table 11 provides a more detailed overview of the number of establishments belonging to one company. For the whole sample, an average of 1.22 establishments is found per company, with a range of 1 to 4,349 establishments. When only looking at multi-site companies, the average number of establishments per company increases to almost four. Finally, as there are some companies for whom a large number of matching establishments were found and who therefore bias this result, the final row shows that an average of almost three establishments was found per company for multi-site companies without the top percentile of observations. Hence, internal labor markets may be analyzed with the data, albeit under the restriction that we cannot fully represent them.

Table 11: Number of linked establishments per company at the company-level

Number of establishments	N	Mean	Std. dev.	Min.	Max.
All companies	535,129	1.22	9.37	1	4,349
Multi-site companies	40,358	3.93	34.01	2	4,349
Multi-site companies (winsorized)*	39,946	2.85	2.43	2	26

Notes: * Companies with a number of establishments in the top 1 percentile have been excluded.
Source: Own calculations based on the record linkage key and the BHP.

6 Representativity of Orbis-ADIAB

We are not able to provide company-establishment matches for the whole universe of establishments in Germany (as recorded by the BHP). As the selectivity of the matches may affect some research questions and analyses, we discuss the representativity of Orbis-ADIAB in the following section.

6.1 Representativity of linked companies

Table 12 provides an overview of the company-level information available in the BvD data, as shown in the row labeled “Orbis” and in the linked company-establishment data (row “Linkage”). The columns describe the number of companies in the original data extract irrespective of the existence of any financial information (1), those with unconsolidated financial information irrespective of the accounting framework (2) and those with financial information according to the GAAP standards (3). Column (4) indicates the number of companies with at least one matched establishment in the BHP and (5) the number of companies in the final dataset fulfilling all the previous requirements. The number of companies in (5) is lower than in (4), as our revisions to improve data quality excluded companies without financial information, as well as those which were part of multiple assignments. Note that the overall number of companies is lower than the overall number of establishments as more than one establishment can be linked to one company.

Table 12: Number of observations at the company-level

	(1) Total	(2) Unconsolidated financial information	(3) Unconsolidated financial information (local GAAP)	(4) BHP	(5) Final Dataset
Orbis	1,948,778	1,058,933	1,058,926		
Linkage	975,880	606,863	606,857	794,938	535,129

Source: Own calculations based on the record linkage key, the BHP and Orbis.

The number of companies in the original BvD data extract of 2014 (N=1,948,778) is reduced by about 50% when considering only linked companies (column 1). The difference between (2) and (3) is negligible and the companies without a local GAAP accounting framework are excluded from the analysis. About 40% of the companies linked in the key do not have corresponding financial information in Orbis and are therefore not included in the final dataset. Comparing (1) and (4) reveals that approximately 80% of the links provided by the key are also included in the BHP. Deducing (2) - (4) from the linked companies in (1) leads to the final number of company-level observations in (5). The reason that the number of observations is

higher in (4) compared to (3) is that (3) and (4) are not consecutive steps. Overall, more than 50% of the links provided in the key can be used for research. The main reasons for the loss in numbers of observations between the key and the final dataset accordingly lie in lacking information for the matches either in the financial data or the BHP.

In a next step, we examine the representativity of our generated research dataset at the company-level. As a basis for the analysis we use the subset of companies in Orbis for which financial data were available. This is the case for 1,058,926 companies as shown in column (3) of Table 12.

Table 13 presents the linkage success rate for the Orbis data. The column “linkage” refers to all companies included in the key (column (1) of Table 12) and the column “final dataset” indicates the number of firms in the final research dataset (column (5) of Table 12). We compare different company characteristics in order to evaluate how selective the final dataset is. The values for Chi^2 indicate that for each group of characteristics under consideration, a significant difference exists between being linked (or not linked) to the average chance of being linked (or not).

We first look at differences between legal forms. As Table 13 indicates, the legal forms “individually owned” and “partnership” are underrepresented compared to other legal forms. This might be due to the fact that the record linkage was not equally successful across different company sizes and that legal forms and company size are correlated. However, we cannot recognize big differences concerning these legal forms between the linkage and the final dataset. Limited partnerships, on the other hand, are slightly overrepresented in both datasets, as are capital corporations and “others”. Overall, we note an unequal distribution of legal forms both in the linkage and the final dataset.

Next, we consider the company size classification according to Orbis to more deeply examine the assumption that larger firms may be overrepresented in the data. We observe that small companies are underrepresented while large companies are clearly overrepresented. We find a difference between the final and the linkage dataset. One possible explanation for that finding is that fewer small companies have unconsolidated accounting reports and are therefore not part of the final dataset.

We further observe whether a company is listed on the stock market. Note that the number of observations for delisted and unlisted groups is rather low, therefore the results concerning these two subgroups have to be regarded cautiously. Listed companies are overrepresented in both dataset. Two explanations account for this overrepresentation. First, listed companies are legally required to report their annual accounts, hence making them less likely to be dropped by one of our data cleaning steps. Second, companies listed at the stock exchange are predominantly large companies and usually have multiple establishments, making them more likely to be found during record linkage. Importantly though, we note few differences between the linked and final datasets.

Table 13: Linkage success rate for companies in Orbis by company characteristics (in %)

	Linkage	Final Dataset
Total	57.29	50.52
Legal form		
Individually-owned company	37.54	34.41
Partnership	33.59	29.97
Limited partnership	57.53	50.69
Capital corporation	70.34	62.98
Other (e.g. registered cooperative, registered association)	62.33	59.15
Chi²	2359.45	1961.54
*Size (BvD)**		
Small company	52.17	44.03
Medium sized company	70.76	67.62
Large company	69.88	66.17
Very large company	72.81	69.31
Chi²	29,844.21	46,715.99
Listed on the stock exchange		
Delisted	74.59	67.62
Listed	79.23	73.63
Unlisted	57.29	50.51
Chi²	187.87	200.28
Federal state		
Schleswig-Holstein	54.1	47.69
Hamburg	49.44	40.49
Lower Saxony	55.99	48.89
Bremen	55.56	48.03
North Rhine-Westphalia	55.33	49.01
Hesse	55.26	48.45
Rhineland-Palatinate	59.12	52.87
Baden-Wuerttemberg	59.42	52.92
Bavaria	55.23	48.06
Saarland	62.9	57.25
Berlin	58.32	50.4
Brandenburg	66.00	59.55
Mecklenburg-Western Pomerania	64.04	57.45
Saxony	69.23	63.04
Saxony-Anhalt	69.44	63.98
Thuringia	68.58	63.03
Chi²	8031.35	9475.51
Status (BvD)		
Active	57.93	51.76
Active (problems)	58.36	45.08

Dissolved	45.29	32.80
Unknown	42.86	42.86
Chi²	3306.62	7791.76
Industry code (first digit of WZ08)		
1 A Agriculture, forestry and fishing	68.78	63.72
2 B Mining and quarrying	62.07	56.60
3 C Manufacturing	76.60	72.10
4 D Electricity, gas, steam and air conditioning supply	26.81	22.92
5 E Water supply; sewerage, waste management and re-mediation activities	70.86	65.81
6 F Construction	74.67	69.73
7 G Wholesale and retail trade; repair of motor vehicles and motorcycles	70.38	63.49
8 H Transportation and storage	73.02	63.27
9 I Accommodation and food service activities	67.82	63.20
10 J Information and communication	68.39	58.56
11 K Financial and insurance activities	37.38	30.18
12 L Real estate activities	38.18	30.05
13 M Professional, scientific and technical activities	36.59	29.00
14 N Administrative and support service activities	62.03	55.68
15 O Public administration and defense; compulsory social security	60.49	55.12
16 P Education	68.57	61.90
17 Q Human health and social work activities	70.44	66.86
18 R Arts, entertainment and recreation	69.92	63.69
19 S Other service activities	63.34	56.01
20 T Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use	74.36	71.79
21 U Activities of extraterritorial organizations and bodies	83.33	66.67
Chi²	127069.4	135000.7
Foundation date		
<=1980	63.63	56.42
1981–1990	63.66	56.26
1991–2000	60.37	53.16
2001–2010	52.50	46.24
>2010	51.30	44.78
Chi²	10293.45	8176.64

Notes: Chi² shows significant differences between being linked or not being linked compared to the average chance of being linked or not linked. The p-values of all Chi² tests are 0.000.
* The variable size was calculated by BvD and does not only indicate the number of employees. Rather, the size for company groups is generated by different steps. In a first step, a primary group is constructed by segments for industries, banks, and insurance companies. Then a ranking by total assets, gross premium and turnover is defined and in step three, a final definition for standard peer groups is made (Bureau van Dijk, 2014).
Source: Own calculations based on the record linkage key and Orbis.

Regarding the federal state we can see that Thuringia, Saxony, Saxony-Anhalt, Saarland, Mecklenburg-Western Pomerania und Brandenburg are overrepresented in both datasets.

Hamburg and other city states are underrepresented. This underrepresentation is probably due to a larger share of small companies in the city states. Note that city states like Hamburg are underrepresented in Orbis, while the opposite is true in the BHP data (as we can see in Table 16). This result is most likely due to the linkage process in which larger companies are predominantly linked.

Concerning the activity status, we do not see large differences between the linked and the final data. As expected, active companies are overrepresented, while inactive ones are underrepresented.

Some interesting patterns emerge for the industry classifications. The industry codes in column (1) of Table 13 indicate that some industries are underrepresented in both the linkage and the final dataset, namely the sectors “electricity, gas, steam and air conditioning supply”, “financial and insurance activities” and “professional, scientific and technical activities”. Other industry sectors are overrepresented in both datasets, namely “manufacturing”, “activities of households as employers; undifferentiated goods- and services-producing activities of households for own use” and “activities of extraterritorial organizations and bodies”. Interestingly, for some industries, the representation differs strongly in the two datasets, namely the sectors “transportation and storage”, “real estate activities”, “information and communication” and “activities of extraterritorial organizations and bodies”. Overall, the distribution of industry sectors remains similar in the linkage and the final dataset.

In terms of the founding year, the table shows an overrepresentation of older companies as companies with a recent foundation year are less likely to be linked. There are different possible reasons for this finding. First, only companies with valid financial information[12] are included in the dataset. Legally, companies can file their financial statements with a time-lag (see Appendix Table A1), which is why young companies are more likely to still be in the filing process within the observation period 2006-2013. Second, older companies are more likely to have expanded their business and founded new branches over time. These companies therefore have, on average, more establishments in Germany, which in turn makes it more likely that at least one of them is found in the linkage process.

Overall the representativity analyses on the company-level demonstrate that companies with specific characteristics are overrepresented. However, most of the observed differences are not unexpected. Nevertheless, one has to keep this selectivity in mind when using Orbis-ADIAB in empirical analyses.

6.2 Representativity of linked establishments

In addition to the representativity checks on the company-level, we also regard the establishment-level. The results provided in Table 15 relate to the establishment-level and are similar to the company-level information provided in Table 12.

Table 15 illustrates that approximately one sixth of the establishments in the BHP data are available in the final research dataset, while about one fourth was linked. The majority of the

[12] This means companies have had to file at least one financial statement between 2006 and 2013 according to the local GAAP on an unconsolidated basis.

losses are due to missing observations in the financial data at the company-level. The decrease in the number of observations from column (3) to column (4) is due to the resolution of multiple assignments.

Table 14: Observations on the establishment-level according to BHP for the years 2006 - 2013

Year	(1) All establishments	(2) Linked establishments	(3) Establishments linked to companies with existing financial data	(4) Final dataset
2006	2,729,252	621,239	433,066	432,674
2007	2,780,754	640,998	455,284	454,874
2008	2,804,470	656,880	475,976	475,550
2009	2,849,632	670,604	498,643	498,185
2010	2,890,234	682,847	523,711	523,239
2011	2,927,181	697,297	543,249	542,771
2012	2,950,533	707,764	544,344	543,863
2013	2,968,382	711,122	534,558	534,077

Source: Own calculations based on the record linkage key, the BHP and Orbis.

Table 16 shows linkage success rates based on establishment characteristics available in the BHP for the year 2013. As the interpretation of the results is analogous to Table 13, only a few insights are highlighted here. First, we observe that linkage success increases with the number of employees in the establishment. This is not surprising, as smaller companies do not have to file annual statements (or only in reduced form according to §267 Handelsgesetzbuch (HGB) – German Commercial Code; see Appendix Table A1 for the requirements for filing an annual statement). Therefore small companies are only represented in Orbis, if they voluntarily filed a financial report. The larger the company, the more extensive is the reporting framework and the more likely it is to be found in Orbis and the more likely we find a corresponding establishment during record linkage. When comparing the two datasets, the pattern remains with the exception of the largest size category, which is underrepresented in the final dataset. Small companies are also underrepresented in the final dataset, again due to the fact that they do not have to report a financial statement.

The analysis further indicates that there is some selectivity with regard to federal state the establishment is located in. While especially the city states Hamburg, Berlin and Bremen are overrepresented, more sparsely populated areas such as Schleswig-Holstein and Rhineland-Palatine are underrepresented. We assume this is due to population density, as well as concentration of companies in specific areas.

A distinct pattern also occurs for the industry codes, whereby industries with a predominance of large companies, such as in manufacturing, are overrepresented, while industries more closely related to services or the public sector (and smaller company sizes) are underrepresented.

Table 15: Linkage success rate BHP by establishment characteristics (in percentages)

	Linkage	Final dataset
Total	23.96	17.99
Employees		
1-10 employees	17.60	12.88
11-50 employees	50.00	39.63
51-100 employees	64.61	49.99
101-250 employees	68.70	50.15
> 250 employees	69.64	47.51
Chi²	318,343.60	249,808.80
Federal state		
Schleswig-Holstein	22.08	16.05
Hamburg	32.28	24.41
Lower Saxony	23.09	17.00
Bremen	30.00	22.05
North Rhine-Westphalia	23.95	17.94
Hesse	24.65	18.24
Rhineland-Palatinate	20.11	14.81
Baden-Wuerttemberg	23.76	18.13
Bavaria	23.50	17.50
Saarland	24.00	18.64
Berlin	27.05	20.58
Brandenburg	24.40	18.93
Mecklenburg-Western Pomerania	22.97	17.74
Saxony	24.49	18.84
Saxony-Anhalt	24.52	18.98
Thuringia	24.44	18.76
Chi²	5,472.28	4,538.66
Industry code (first digit of WZ08)		
1 A Agriculture, forestry and fishing	9.20	6.81
2 B Mining and quarrying	62.97	50.19
3 C Manufacturing	49.17	40.00
4 D Electricity, gas, steam and air conditioning supply	61.24	53.03
5 E Water supply; sewerage, waste management and remediation activities	47.38	38.00
6 F Construction	33.47	27.59
7 G Wholesale and retail trade; repair of motor vehicles and motorcycles	37.84	25.26
8 H Transportation and storage	36.58	28.37
9 I Accommodation and food service activities	10.90	7.60
10 J Information and communication	54.41	42.71
11 K Financial and insurance activities	26.51	13.83
12 L Real estate activities	15.01	12.12
13 M Professional, scientific and technical activities	28.98	23.22

14 N Administrative and support service activities	31.94	23.67
15 O Public administration and defense; compulsory social security	3.71	3.15
16 P Education	12.32	9.06
17 Q Human health and social work activities	7.15	5.78
18 R Arts, entertainment and recreation	17.77	13.70
19 S Other service activities	7.79	6.16
20 T Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use	0.56	0.35
Chi²	403,745.1	296,016.6
Foundation date		
<=1980	30.61	22.64
1981–1990	26.38	21.43
1991–2000	24.36	20.24
2001–2010	22.92	18.52
>2010	20.43	8.99
Chi²	12,889.04	36,998.16
Median income missing		
Not missing	36.31	27.75
Missing	9.26	6.38
Chi²	295,857.70	227,801.10

Notes: Chi² shows significant differences between being linked or not being linked compared to the average chance of being linked or not linked. The p-values of all Chi² tests are 0.000.
Source: Own calculations based on the record linkage key and the BHP.

Compared to Orbis, a new aspect can be analyzed with the BHP, i.e. the wage structure of establishments. We find, that there is a slight underrepresentation of establishments with missing values for the median wage information. This finding is due to the fact that such missing values only occur for very small establishments as they are more likely to have no employees working full-time. Only the latter are used to compute the median wage variable on the establishment level.

Finally we evaluate mean differences between the BHP and the final linked dataset in order to detect significant differences. Table 17 shows mean differences and corresponding t-values between the linked dataset and the BHP for the year 2013. Noteworthy is the underrepresentation of women and the overrepresentation of full-time workers in the final dataset. We assume a correlation, as women often work in part-time jobs. While unskilled or semiskilled tasks are underrepresented, more highly skilled tasks are overrepresented in the final dataset. Thus, it seems that a slight selection according to skill level may exist. The mean difference in the median wage of all full-time employees probably correlates with the overrepresentation of higher skilled workers and the overrepresentation of larger companies.

Table 16: Mean differences between BHP and the final dataset at the establishment-level, t-test of difference

	(1) Final dataset	(2) BHP	(3) Difference	(4) t-value	(5) Number of observations
No. employees	27.18	8.61	18.57	126.59	2,968,382
Share female employees	0.44	0.63	-0.19	-323.70	2,968,382
Share trainees/apprentices	0.03	0.02	0.01	71.09	2,968,382
Share full-time	0.55	0.28	0.27	506.36	2,968,382
Share part-time	0.16	0.20	-0.04	-96.77	2,968,382
Share fixed term employees	0.10	0.07	0.03	106.65	2,968,382
Share Germans	0.92	0.89	0.03	81.15	2,968,382
Share employees unskilled/semiskilled tasks	0.18	0.22	-0.04	-83.05	2,968,382
Share employees skilled tasks	0.59	0.55	0.04	68.04	2,968,382
Share employees complex tasks	0.13	0.07	0.06	174.49	2,968,382
Share employees highly complex tasks	0.10	0.05	0.05	192.74	2,968,382
Mean age of the total of employees	43.00	44.77	-1.78	-110.58	2,968,382
Median wage all full-time employees	86.51	70.95	15.55	260.36	1,613,009

Note: Comparisons of means at the establishment-level using the BHP for the year 2013. Results of t-tests provided in the column (4). The number of observations is lower for the median wage of all full-time employees because only establishments with at least one full-time employee are used to calculate the median wage (see Schmucker et al., 2016).
Source: Own calculations based on the record linkage key and the BHP.

Overall, as with the company-level analysis, we find some selectivity of the linked establishments compared to all establishments in Germany in 2013. However, the differences are not large and can mostly be explained by the establishment's characteristics. Overall, while researchers should be aware of potential selectivity, we do not conclude that large representativity issues exist.

7 Conclusion

We describe the steps taken to create a research dataset from the key developed through record linkage by Schild (2016). We show that merely creating a correspondence table, i.e. a key between companies and their dependent subunits, does not suffice to provide a useful research dataset. Thus, several additional steps have to be taken to develop a research dataset. These steps include, for example, dealing with false links, such as establishments being assigned to more than one company, or assigning headquarters. We then assess the representativity of the research dataset and conclude that while some selectivity issues may exist, we are confident that we can explain these sufficiently to allow researchers to take account of them.

Remaining challenges will be addressed in an internal test phase of the research data and result in subsequent revisions following feedback loops. Within this test phase, we will consider how to draw useful samples of the data to be provided to researchers and how to anonymize the data to conform with data protection legislation. Furthermore, we expect that the test phase will reveal any potential data quality issues and help with data documentation.

What we learned in these analyses, e.g. about characteristics of companies that are underrepresented in the linked data, will also help us improve the linkage key. In addition to the Orbis extract of 2014, which was used to create the existing linkage key, the improved and extended linkage key will also rely on Orbis extracts for the year 2015 – 2017. This work is currently ongoing.

References

Antoni, M., Schnell, R. (2017). The Past, Present and Future of the German Record Linkage Center (GRLC). In: Journal of Economics and Statistics, online first.

Biewen, E., Gruhl, A., Gürke, C., Hethey-Maier, T., Weiß, E. (2012). “Combined Firm Data for Germany” — Possibilities and Consequences of Merging Firm Data from Different Data Producers. In: Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, 132(3), 361–377.

Bureau van Dijk (2014). Dafne – User guide. Retrieved from: https://webhelp.bvdep.com/Robo/BIN/Robo.dll?project=84_EN&newsess=1&refer=https%3A//webhelp.bvdep.com/robo/projects/wh00000/UG.htm. Last accessed December 07, 2016.

Fischer, G., Janik, F., Müller, D., Schmucker, A. (2009). The IAB Establishment Panel * things users should know. In: Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, 129 (1), 133–148.

Herzog, T. N., Scheuren, F.J., Winkler, W.E. (2007). Data Quality and Record Linkage Techniques. New York: Springer Science + Business Media.

Schäffler, J. (2014). ReLOC linkage: a new method for linking company-level data with the establishment-level data of the IAB. FDZ-Methodenreport, 05/2014 (en).

Schild, C.-J. (2016). Linking “Orbis” Company Data with Establishment Data from the German Federal Employment Agency. German RLC Working Paper, wp-grlc-2016-02.

Schmucker, A., Seth, S., Ludsteck, J., Eberle, J., Ganzer, A. (2016): Establishment History Panel 1975-2014. FDZ-Datenreport, 03/2016 (en).

Appendix

Table A1: Legal requirements for filing an account

	Small company	Medium company	Large company
Legal reference	§ 267 HGB	§ 267 HGB	§ 267 HGB
Number of employees	1–50	51–250	250+
Turnover in 1000 €	Less than 9.681	9.681–38.500	38.500+
Total assets in 1000€	Less than 4.841	4.841–19.250	19.250+
Requirements		at least two criteria must apply	
Required announcement	- balance sheet - notes on the accounts	- balance sheet, statement of income - notes on the accounts	- balance sheet - statement of income - notes on the accounts
Filing location	Register Court	Register Court	Bundesanzeiger
Maximum filing duration after the end of the year		12 months	
Average filing time		9–12 months except for quoted companies	

Notes: These firm sizes do not correspond to the firm sizes in Table 13.
Source: Handelsgesetzbuch (HGB).

Imprint

FDZ-Methodenreport 4/2018 (EN)

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Dana Müller, Dagmar Theune

Technical production
Dagmar Theune

Download
http://doku.iab.de/fdz/reporte/2018/MR_04-18_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:
Manfred Antoni
Institute for Employment Research (IAB)
Research Data Centre (FDZ)
Email: Manfred.Antoni@iab.de

Marie-Christine Laible
Institute for Employment Research (IAB)
Research Data Centre (FDZ)
Email: Marie-Christine.Laible@iab.de