German Record Linkage Center
WORKING PAPER SERIES
NO. WP-GRLC-2016-02 | NOVEMBER 30, 2016

Linking "Orbis" Company Data with Establishment Data from the German Federal Employment Agency

Christopher-Johannes Schild

Linking “Orbis” Company Data with Establishment Data from the German Federal Employment Agency

Christopher-Johannes Schild

German Record Linkage Center: Working Paper

Contents

Abstract 1

1 Introduction 2

2 Brief Description of both Datasets 2

3 Selection of Identifiers 3

4 Data Cleaning 6

5 Matching Procedures 6

6 Matching Result 13

7 Summary 17

A Further Tables 18

References 19

Abstract

We perform a record linkage of company data from the database Orbis of the data provider Bureau van Dijk with establishment data from the Institute for Employment Research (IAB) at the German Federal Employment Agency. We describe both data sets with regard to the identifiers available, as well as the methods that were used for this record linkage. For 82.4% of Orbis companies with more than 5 employees, at least one IAB establishment could be assigned. We perform a series of tests to verify the overall quality of the record linkage.

Keywords: company data, establishment data, administrative data, record linkage, entity resolution

Acknowledgements: The author would like to thank Manfred Antoni, Stefan Bender, Johanna Eberle and Marie-Christine Laible for helpful comments.

1 Introduction

In this paper, we describe the first comprehensive attempt to link almost the entirety of companies registered in the German trade register, plus an additional set of companies included in the Orbis-database of the data provider Bureau van Dijk, to data on establishments from the Institute for Employment Research (IAB) of the German Federal Employment Agency (BA). The linked dataset provides the information about which IAB establishments belong to the same company. It also enables, for example, joint analyses of financial information for German companies with administrative labor market data of the IAB.

In the next section, both datasets to be linked are briefly described. In the third section, we discuss the suitability of potential identifiers that can be used for assigning establishments to companies when no matching key exists yet. In the fourth section, we describe data cleaning procedures performed on the raw data. In the fifth section, the applied record linkage procedures are outlined. In the sixth section, the result of the linkage is described and interpreted.

2 Brief Description of both Datasets

Bureau van Dijk (BvD) is a commercial provider of firm data and business intelligence. Their database "Orbis" contains business records for 1,938,990 firms (as of January 2014), of which 1,627,668 were marked as active (`https://orbis.bvdinfo.com`). The subset of the variables included in the data extract available to the IAB comprises unconsolidated financial data. A detailed list of the variables in the IAB's data extract can be found in Antoni et al. (forthcoming).

The administrative research data of the Institute for Employment Research (IAB) contain detailed information on the employment history of all employees liable to social security contributions, on marginal part-time employment, benefit recipients, registered job-seekers and participants in programs of active labour market policies on a daily basis. For all the employees liable to social security contributions, a numerical establishment identifier is available, which makes it possible to aggregate to the establishment level, generating an establishment dataset for all establishments with at least one employee. This establishment data cover the years 1975-2010 and are available as a research dataset, the "IAB Establishment History Panel (BHP)" [1] (Gruhl et al., 2012).[2] The research data on establishments contain roughly 2.7 mio active establishments of the BHP.

The resulting joined company-establishment dataset naturally both contains variables on the establishment level from the IAB and on the firm level from the Orbis database. The available establishment variables are those of the BHP dataset. The company level variables are available from the Orbis database (Antoni et al., forthcoming).

[1] At the time of linkage this was the current version of the BHP.
[2] Due to data protection regulations, this research data do not contain firm names or individual names.

Companies do not necessarily have to own establishments according to the IAB definition: for example, a small restaurant in which family members are the sole labor providers may not have any employees liable to social security contributions. Even an incorporated company may not have any employees liable to social security contributions, for example if this company only employs temporary agency workers, who, while they are of course also liable to social security contributions, are registered for a temporary work agency. From the establishment perspective, while it is clear that all establishments have to be owned by some legal entity, i.e. usually a private company, this entity does not need to be in the Orbis data; the reason for this is that the Orbis companies are (roughly) identical to the trade register companies, and requirements to register are conditional on legal form and size.[3]

It is important to note for this linkage project that we do not know how many companies in the Orbis data actually own an establishment according to the IAB definition. Neither do we know the share of IAB establishments for which the owning company is in the trade register, i.e. the Orbis data. We have to keep this in mind when interpreting any measures of the goodness of our match, such as the overall share of companies for which we find at least one establishment (see section 6).

3 Selection of Identifiers

The problem to be solved with this project was to assign to each company in the Orbis data, that in fact has at least one establishment according to the IAB definition (i.e. at least one employee liable to social security contributions), every establishment in the IAB data that belongs to this company ("1:m match"). Since there was no common identifier available to make such an assignment, the record linkage had to be conducted using alternative identifiers, such as establishment names and addresses. To this end, establishment names and addresses, that due to data privacy reasons are not included in the IAB establishment research data, were temporarily acquired from the data warehouse of the Federal Employment Agency, and temporarily linked to the IAB establishment data, for the purpose of the record linkage.[4]

The variables considered for solving this linkage problem are the company / establishment name, legal form[5], address (city, postal code, street, housenumber), number of employees and main industrial sectors of activity. These identifiers are described in more detail below, along with a discussion of their suitability for performing a record linkage of firms and establishments.

[3] Incorporated companies such as AG, GmbH, OHG are required to register, no matter how large. But non-incorporated companies, such as sole-proprietorships, have to register only if they exceed certain size thresholds that are set by the chambers of commerce (a usual rule of thumb is 500,000 Euros of revenue).

[4] We did not attempt to use establishment name and address information older than 5 years.

[5] The legal form of the company is included in the Orbis data. For establishments, the legal form of the owning company was extracted from the establishment name.

Company Name and Legal Form

The Orbis data include the name of the company as recorded in the German trade register. The registered company name has to fulfill the following requirements (IHK Köln (2012), cited after Schäffler (2014)). The company name ...

- can be “any family name, term or any freely chosen name and may consist of several words”
- “has to comprise the legal form of the company”
- “may include company slogans”
- has to be suitable to “identify the registering trademan” and has to have “discriminatory power”

Discriminatory power can be achieved by adding further name components, or adding a city name to the company name, such as adding the family name “Schmidt” and the city name “Frankfurt” to the insufficiently discriminatory name “Immobilien GmbH”, which would create “Schmidt Immobilien Frankfurt GmbH”. The establishment name and address data that was temporarily used to enhance the IAB establishment data for the purpose of the record linkage include the name of the company that owns the establishment. This data is generated by the office of the Federal Employment Agency that assigns establishment identifier numbers: this office receives the company name in written form and inputs the name in three textfields with 30 characters each. Since this entering is done manually, errors and irregularities arise, such as typing errors or irregular occurrences of extra name components such as names of owners or company slogans.

A high discriminatory power of company names in combination with the legal form makes this combination a preferred identifier. Since for companies with multiple establishments, such as is often the case with companies in the retail sector and larger companies, the identifier combination has the additional advantage that it is independent of the geographic location. In spite of the legal requirement that company names have to have discriminatory power, for roughly 10% of companies in the Orbis data there is at least one twin in terms of company name / legal form, which may be explained by the fact that identical names may be tolerated when the field of business, geographic area of business, or both differ sufficiently, in order to assure that the companies can be distinguished from each other. Discriminatory power of names and legal form is therefore imperfect without information on the geographic area of business (which is not available in either the company data or the establishment data), which means that for those companies, it is a priori clear that it will be difficult to distinguish true from false matches.[6]

[6] It may be possible to assign establishments to these companies, in principle, by including geo-information on companies’ and establishments’ addresses. However, this will not enable us to find establishments that are located within the geographic area of business, but not located at the exact same place as the company headquarters. Such a procedure is therefore likely to lead to an oversampling of one-establishment companies and of main establishments of multi-establishment companies (see below).

Location of Companies and Establishments

The Orbis company data contain addresses of companies, but no information on the area of business.[7] The IAB establishment data was temporarily enhanced by establishment addresses for the purpose of the record linkage (see above). Since multi-establishment companies may comprise establishments that are dispersed over a region or even the whole country, using information on company and establishment location is likely to lead to an overrepresentation of one-establishment companies and of main establishments of multi-establishment companies in the final linked dataset, i.e. to an "oversampling" of such companies in the final matched dataset. On the other hand, the likelihood of correctly assigning establishments to single-establishment companies increases dramatically once we make use of addresses, since including addresses in the matching enables us to lower the minimum quality score of an error-tolerant comparison of names, while holding the chance of false assignments constant. So we assume that we face a trade-off between oversampling of single-establishment companies on the one hand, and increasing the matching rate on the other hand. Assuming this trade-off, we decided that to a certain extent, addresses should be used. We try to limit the oversampling of one-establishment companies and main establishments caused by using addresses by putting matching methods that use addresses last in the sequence of matching steps.

There is at least one specific case in which the problem of oversampling of single establishment companies may be less severe: in the case of the legal form "registered merchant" ("Eingetragener Kaufmann"), one can argue that the vast majority of these companies only has one establishment, which then allows us to use addresses as an additional identifier. But even with German companies that have a different legal form, such as the most common "GmbH",[8] we know from empirical evidence that the vast majority only has one establishment.

Main Sector of Activity

Both datasets contain information on the main sector of activity. The suitability of this variable is limited for two reasons: first, for larger, multi-establishment companies, using this identifier may lead to an oversampling of establishments that are active in the main sector of the company, and to an undersampling of establishments "untypical" for the company. Generally, it may lead to an oversampling of one-establishment-companies. Secondly, the main field of activity for the IAB establishments is based on employment, while for the Orbis data it is based on revenue. Nevertheless, the industry code may be an additional means to differentiate between companies in the Orbis data that have an identical name, such as "Fischer GmbH", which is why we decided to use at least the first digit of the industry code[9] as an additional identifier in later matching steps.

---

[7] Except for very few companies.

[8] "**G**esellschaft **m**it **b**eschränkter **H**aftung", a limited liability company according the German limited liability company law ("GmbH-Gesetz")

[9] The German Classification of Economic Activities, Edition 2008 (WZ2008) was used.

As another potential identifier, one could consider the number of employees. However, there are at least two problems that led us to refrain from using this variable: first, this variable's quality is certain to be limited, since "employees" in the sense of the IAB data (employees liable to social security contributions), is not the same as how a BvD clerk making the entry in the Orbis database may understand it. This is true because the colloquial concept of what constitutes an employee may also comprise working proprietors, family workers, or temp workers, who in the IAB data are in fact registered with a temporary work agency. Secondly, we do not know a priori how many establishments belong to each company, which would limit the function of employment as an identifier to the case where aggregated establishment employment exceeds company employment.

4 Data Cleaning

Data cleaning, or "preprocessing" of raw data, is an essential step before employing similarity algorithms in record linkage (Herzog et al., 2007; Schnell et al., 2003). Preprocessing means removing spelling mistakes and known variations in correct notations (i.e. abbreviations etc.), thus equalizing differing entries which are known to refer to the same object ("standardization"), extracting variables from common text fields ("parsing") and eliminating implausible values, such as eliminating negative values for the number of employees ("plausibility-based elimination") (Herzog et al., 2007).

Standardization involved steps common to all string variables (replacing German Umlauts, removing leading and trailing blanks; see Schnell et al., 2003). Company names were parsed into subcomponents by using separating characters such as spaces and hyphens, then concatenated in different combinations of name components to be used as identifiers (see section 5). For place names, common spelling mistakes, abbreviations and inconsistently used geographic name complements (such as "Frankfurt am Main" or "Frankfurt (Main)") were collected and corrected. With regard to street names, the common name component "STRASSE" was standardized to its common abbreviation "STR" ("stemming"). Typical spelling mistakes of streets named after famous persons were collected and corrected. Any numbers contained in street names that could be identified as certainly being a component of the street name were spelled out (e.g., "STR DES 17. JUNI" became "STR DES SIEBZEHNTEN JUNI"), in order to increase chances of correctly parsing street names and house numbers. As a final preprocessing step, all letters were capitalized.

5 Matching Procedures

Depending on data quality, the size of the data sets to be linked, and the available identifiers, different matching methods may be optimal. Our choice of method was guided by best practice insights gained from previous comparable projects at the IAB, as well as extensive pre-tests (by clerical review) of linkage methods for this project. Below we first describe the employed linkage methods generally. Then we briefly describe each linkage method as applied in the final version of the linkage process for this project.

5.1 General Description of Employed Matching Methods

Deterministic matching

With deterministic matching, or "exact matching", both records have to share the exact same values for the complete set of available identifiers (Herzog et al., 2007).

Distance-based matching

Distance-based matching can be used when a record's identifier values contain noise, such as spelling mistakes, since in these cases deterministic linkage will generate false negatives[10]. For comparing error-prone string identifiers, string comparator algorithms are employed (for an overview see Herzog et al., 2007). Among these algorithms, **Jaro Metrics** are particularly suited to capture typical human typesetting mistakes, since they emphasize transposition of characters, i.e. switching of single character positions. The **Jaro-Winkler** variant of this metric gives more weight to initial characters of strings, which can be useful if the likelihood of transpositions is lower for the first characters of a string. That is typically the case with individual names (Herzog et al., 2007). However, in the case of company names, which are concatenations of single components (such as family names, activity descriptions and others), and which are often characterized by switched positions of these components, the use of **n-Grams** has proven to be a more suitable string comparison algorithm, since they are insensitive towards the position of an n-series of characters. A string of length m has $m - n + 1$ n-grams. For example, the name "MERCEDES" has a length of $m = 8$ and 7 substrings of length $n = 2$ ("bi-grams"): "ME", "ER", "RC", "CE", "ED", "DE", and "ES". With n-grams, string similarity measures can be constructed by counting the number of common substrings and dividing by either number of n-grams in the shorter string (Overlap coefficient) or the longer string (Jaccard similarity) or by the average number of n-grams of both strings (Dice coefficient). The fact that n-grams do not consider the order of string sequences can be an advantage when strings are expected to consist of substrings with several likely possibilities to arrange the substrings, as often the case with company names such as in "Siemens Healthcare - Customer Solutions" vs. "Customer Solutions, Siemens Healthcare".

Identifier specific comparison criteria

There are however disadvantages of n-grams as compared to exact similarity: one are large computational costs, the other are sensitivity towards insertions, abbreviations, suffixes and prefixes, and also the insensitivity towards positions of substrings. For the latter problem, consider the example "BMW Bayerische Motorenwerke" vs. "Bayerische Motorenwerke Niederlassung Maisach" or vs. "BMW Niederlassung Maisach". The above example

[10] A "false negative" is a pair of an establishment and a company that is classified as a non-match, even though the pair really is a match.

points to the possibility that it can be better to complement n-grams with identifier specific, theoretically derived comparison rules from what we know about typical construction rules for these specific identifiers. For company names in general, and for the BMW example above, consider the following two rules: “first 15 characters identical, optionally switched to the right by up to 4 characters” and “first 3 characters identical”. Extreme examples for the suboptimality of a sole reliance on n-grams can arise if discriminatory power is carried by a single character, as in “BKG Immobilienverwaltung” vs. “BKB Immobilienverwaltung”. Note how in this example, the n-gram score is high, even though both records are very likely not a match.

Prediction based on subsample regressions

Optimal selection and weights of identifier comparison rules can be achieved by supervised machine learning approaches. This is conceptually and computationally cumbersome, and an alternative is to theoretically derive identifier specific comparison rules, such as the “rule of first three letters” mentioned above, and combine them with n-gram and other more general string comparison algorithms. A pragmatic approach to increase both matching rates and matching precision is then to first use n-grams deliberately, to fetch a large number of possible matches, with a likely large share of false positives. Then, secondly, manually classify a (large) random subsample of assigned matches into true and false positives. In a third step, the set of theoretically derived identifier specific string comparators can then be regressed on “true match” for this subsample and the regressors adapted for best fit. If the fit of the prediction model is sufficiently good, this model can then be used to predict the likelihood of a match for each possible match identified by the n-gram (or other) algorithm in the previous step.

Array matching

An array match means comparing all representations of an identifier in the one file with all representations of that identifier in the other file, and to assign the highest similarity value of all these comparisons to the record pair. This is a suitable strategy when there are several identifier variables in at least one data set, that may equal the value contained in the variable of the other data set. This is the case when, for instance, the first data set comprises the variables “last name” and “maiden name”, and the second data set comprises only “last name”, and there is no information on the marriage day or current marital status of that individual.

Blocking

Comparing millions of company names with millions of establishment names by string comparison algorithms will result in a total number of comparisons in the order of several trillion (the cross product of both name vectors), resulting in prohibitively long calculation times.

Blocking is a very effective way of reducing calculation duration. Traditional blocking involves restricting comparisons to record pairs with exact similarities on one or more identifiers, such as postal code, which can drastically reduce the number of comparisons. Since exact blocking excludes the possibility of finding matches of individuals with erroneous or missing values in the blocking variables in one data set, this can lead to false negatives. Therefore, it is advisable to use different blocking variables in subsequent steps, or different combinations thereof.

5.2 Project Specific Matching Strategies

A general consideration for this project was to find an optimal balance between a) avoiding false positive assignments and b) avoiding the use of identifiers that lead to a systematically higher likelihood of finding establishments for companies with certain properties, such as identifiers based on addresses.

Rule of trade register uniqueness

A convenient specificity of the Orbis data consists in the fact that this data includes practically all German firms that are registered in the German trade register. This is a useful property of this dataset, not only because it guarantees a large number of cases, but also because it makes it possible to evaluate for certain identifiers or combinations of identifiers, whether they are unique in Germany. In order to avoid false and multiple assignments of establishments to firms, one can argue that it should only be attempted to match on identifier combinations that are unique in the Orbis data (and thus in the German trade register). For example, if “Lautenfeller GmbH” only exists once in the Orbis data, this provides us with some confidence that this identifier combination is in fact sufficiently rare to assume all establishments in Germany with a similar name and legal form to belong to this company. As another example, note that this rule prohibits us to match any establishments with the name “Fischer GmbH”, which occurs more than 50 times in the Orbis data, solely by using the identifiers name and legal form. However, we may be able to correctly assign them if we include further identifiers, such as the industry code. We argue that the reduced matching rate that results from including only unique identifier combinations is unproblematic as long as one assumes that having a common, less identifiable firm name is not systematically related to relevant firm characteristics.

Limiting computational costs

With each linkage step, each establishment identification number for which a company could be assigned was erased from the IAB-Establishment-File, reducing the number of unmatched cases with each further step. This procedure aimed at reducing the necessary number of pairwise comparisons for each successive linkage step, thus limiting computational costs. Another measure to reduce computational costs was to substitute n-gram

comparisons with exact comparisons of parsed and rearranged name components. For example, before it was tried to compare “Siemens Healthcare - Customer Solutions” in the one dataset with “Customer Solutions, Siemens Healthcare” in the other dataset by the use of n-grams, after preprocessing, the latter string was parsed into “CUSTOMER”, “SOLUTIONS”, “SIEMENS” and “HEALTHCARE” and then concatenated and combined to name component triplets without changing the order (“CUSTOMERSOLUTIONSSIEMENS”, “CUSTOMERSOLUTIONSHEALTHCARE”, “CUSTOMERSIEMENSHEALTHCARE” etc). The former string was likewise varied to all possible triplets of name components, but with including all possible variations of the name components’ order. Then all such generated name variations were compared for exact identity.[11] Lastly, the number of pairwise n-gram comparisons was limited by blocking over name components and geo identifiers such as postal codes. To give an example for blocking over name components, “BMW BAYERISCHE MOTORENWERKE” would only be compared with those names that also contain either the parsed substring “BAYERISCHE”, or “BMW”, or “MOTORENWERKE”, thus tremendously limiting the number of pairwise n-gram comparisons, without too much risk of not finding true positives (“false negatives”), since only one name component has to be exactly identical in order for an inexact string comparison to take place.

Regarding the technical infrastructure, deterministic and rule-based[12] matching was done with Stata, for the distance-based[13] matching the software “Merge Toolbox (MTB)” was used.[14] Preprocessing and linkage calculations were done on a windows server system with 48 cores and 128GB ram.

Specific matching steps

We started with linkage by exact agreement on the preprocessed firm / establishment name and exact identity of the legal form that was extracted from the firm name field. For this exact linkage step, as for all other linkage steps, we only considered combinations of identifier values that were unique in the Orbis data. The order of all following linkage steps was determined by successively a) relaxing identity rules (thus risking more false positives) and b) trying to limit the increase in falsely positive assignments by (cautiously) making use of additional, non-name and non-legal form identifiers.

In the final version of the linkage, 17 matching steps were performed:

1. exact long name[15] and legal form
2. exact short name[16] and legal form

[11] While regarding the rule of trade register uniqueness of identifier combinations (see above). Note that the use of component triplets also reduces sensitivity towards insertions, prefixes etc.
[12] Such as “rule of first three letters”, see below.
[13] Such as n-grams.
[14] See Schnell (2004) for details on the MTB as well as `http://record-linkage.de`.
[15] “Long name” refers to all name components that do not describe the legal form, concatenated to one (long) name.
[16] “Short name” refers to all name components that occur before the first name component that describes the

3. exact long name and first 4 digits of the postal code (only sole proprietorships)
4. exact short name and first 4 digits of the postal code (only sole proprietorships),
5. n-grams of names (array) and exact legal form
6. exact long name and legal form and first digit of the industry code
7. exact short name and legal form and first digit of the industry code
8. exact long name and place
9. exact short name and place
10. name component triplets and first 4 digits of the postal code
11. name component triplets and first 3 digits of the postal code
12. name component triplets and place,
13. n-gram names (array) and exact first 4 digits of the postal code,
14. exact long name w/o activity components and first 4 digits of the postal code,
15. exact short name w/o activity components and first 4 digits of the postal code,
16. exact long name w/o activity components and first 3 digits of the postal code
17. exact name w/o activity components and first 3 digits of the postal code

Steps (1) through (4) rely on exact identity of name and legal form, only resorting to addresses in the case of sole proprietorships. Step (5) attempts to find further establishments with error-tolerant string comparisons, without resorting to geo identifiers. Steps (6) and (7) aim to find establishments for firms without a unique combination of name and legal form, by taking the first digit of the industry code as an additional distinguishing property.[17] Steps (8) and (9) take addresses for all legal forms as an additional identifier. Steps (10) to (13) are n-gram or less computationally intensive n-gram-like comparison rules, making also use of addresses in order to decrease risk of falsely positive assignments (at the risk of oversampling single establishment companies). Steps (14) to (17) are variants of steps (3) and (4), with all legal forms included and with the additional property that very common name components that describe the activity of a firm are removed.[18]

legal form, concatenated to one name. This means that “short name” is identical to “long name”, except that all name components that occur after the legal form are discarded.

[17] The industry code, WZ2008, actually has 5 digits, however it was not attempted to use more than the first digit for this strategy, both due to quality concerns regarding the variable and due to the problem of further increasing the risk of oversampling establishments of single establishment companies or of establishments of multi-establishment companies that are active in the main industry sector of the firm.

[18] Take the above example: “BKG Immobilienverwaltung” and “BKB Immobilienverwaltung”. Removing components that describe the activity, “Immobilienverwaltung” (real estate management), these firm names would become “BKG” and “BKB”, respectively.

Prediction and elimination of false positives by subsample regressions

In order to reduce the risk of falsely positive assignments by inexact matching methods, we took a random subsample of 5000 pairs of potential matches (company name and establishment name) gained by more problematic n-gram or n-gram-like matching methods (5), (10) to (13), as well as (14) to (17), and classified each of these 5000 matches manually as either a true or a false positive. In the course of this classification, we looked for recurring characteristics that distinguished true positives from false positives and derived the following rules from the data (for some of which there is a clear theoretical justification):

a) the first three character substring of one string shows up anywhere in the other string, but not at the beginning of the other string (company short forms or initials, such as "BMW", are likely to be positioned at the beginning and have a high information content and should be looked for anywhere in the other string)

b) character positions 1-5 of one string are equal to character positions 2-6 or 3-7 or 4-8 (spaces removed) of the other string (irregularly used short or medium length prefixes, mostly company short forms or initials (often three letters), first names of company owners)

c) the last 6 character substring of one string show up anywhere in the other string, but not at the end of the other string (shifted name components)

d) exactly one character in the one string is not included or replaced by a different character in the other string

e) both strings share a substring of length 6 that in one of both strings must be positioned around the middle of this string

f) character positions 1-3 of one string are equal to character positions 2-4 (spaces removed) of the other string (short prefixes such as owner initials may shift company initials (often three letters) to the right)

g) the last 6 character substring of one string is identical either to the last 6 character substring (-6 to -1) of the other string or to characters -7 to -2 (up to -9 to -4) (short suffixes such as unrecognized legal form indicators)

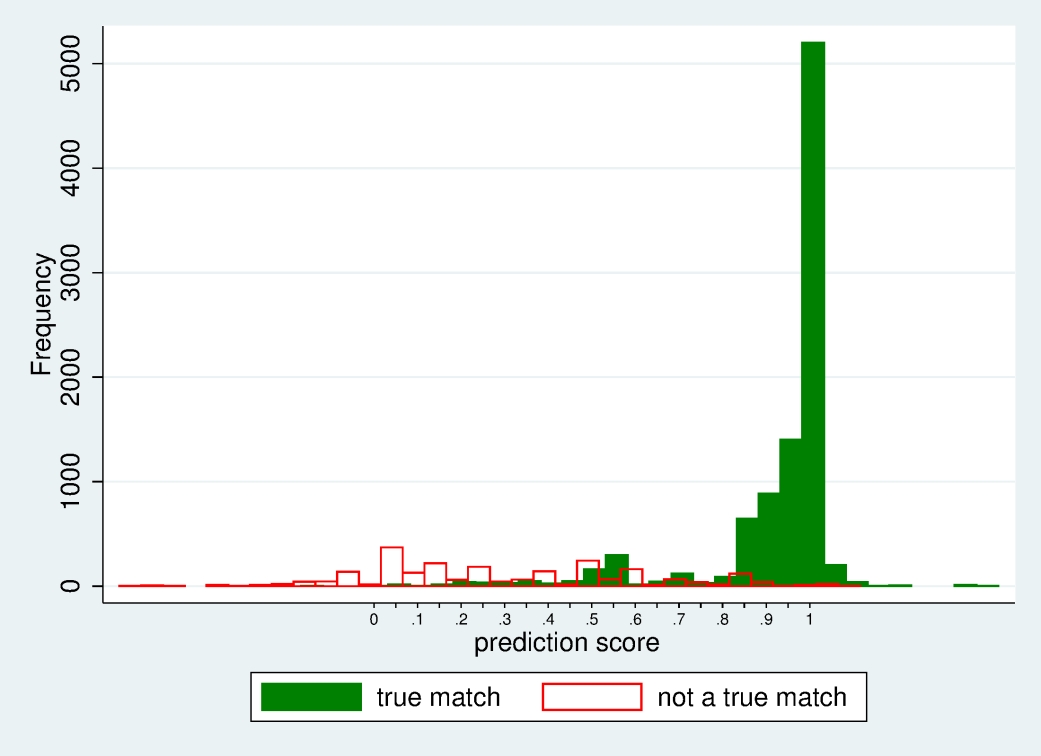
h) the first 15 characters substring is identical, optionally shifted to the right by one character

i) the last 15 characters substring is identical, optionally shifted to the left by up to three characters

The rules were calculated for all matches that were generated by the above mentioned problematic matching methods. The above rules where then regressed on the dummy "true positive (0/1)" for the 5000 subsample, including interaction terms of all rules with the matching method with which each match was found. The linear regression model yielded

an R^2 of 0.66. The model was then used to calculate a prediction for the entire sample. Figure 1 shows the distribution of true and false matches for the subsample of 5000:[19]

Figure 1: Prediction vs. true classification of matches for a random subsample



Based on the result of the subsample regression of firm name specific string comparison rules on true positive assignment, it was decided to choose a cutoff quality score of 0.75. About 27,000 matches were reclassified to non-matched by this procedure (see No 19 in table 4 in section A).[20]

6 Matching Result

Since we do not know how many companies in the Orbis data actually own an establishment in the IAB definition,[21] quality indicators, such as the overall share of companies to which at least one establishment can be assigned, are of limited value for assessing the overall matching success. Luckily, for 571,662 out of 1,627,668 companies marked

[19] Note that the potential matches shown in the graph add up to more than 5000, this is due to multiple occurrences of potential match name pairs for multi-establishment companies.

[20] Table 4 shows for each company with at least one successfully assigned establishment the matching method that led to this assignment (in the case of exactly one assignment) or the best matching method among all assigned establishments. It is noteworthy that for most of the companies (87.8% of all companies matched), at least one of the establishments matched was assigned by exact comparisons and for only 12.2% the best matching method was through inexact comparison methods. Note that since table 4 only shows the matching method of the best matched establishment for each company, this 12.2% understates the total share of matches that was found through inexact matching methods.

[21] Again, note that companies need to have at least one employee liable to social security contributions in order to be in the IAB establishment data.

as “active” in January 2014 in the Orbis data, total company employment[22] is available. This variable (EMPL) likely does not accurately measure the number of employees liable to social security contributions, since it may also include working proprietors, working family members, etc. (see the discussion in section 3). However, we can assume that only a small share of companies with more than a handful (say, companies with EMPL larger than five) employees does *not* have at least one regular employee, and that therefore, for most of these companies[23] there should be at least one corresponding establishment in the IAB data.

Table 1: Match Success for Companies Active 2014, by EMPL Size Class

	no (=0) or at least one (=1) establishment assigned					
Size Class (EMPL)	**0**		**1**		**Total**	
	%	N	%	N	%	N
1-5	44.5	140,868	55.5	175,846	100.0	316,714
6-10	20.0	16,296	80.0	65,268	100.0	81,564
11-25	17.8	15,677	82.2	72,583	100.0	88,260
26-50	16.2	6,794	83.8	35,046	100.0	41,840
51-100	15.2	3,476	84.8	19,331	100.0	22,807
101-250	13.2	1,820	86.8	12,013	100.0	13,833
250+	12.5	829	87.5	5,815	100.0	6,644
missing	56.4	595,374	43.6	460,638	100.0	1,056,012
Total	48.0	781,134	52.0	846,540	100.0	1,627,674

Source: Own calculations based on Orbis company data and IAB establishment data.

Table 1 shows matching success rates, i.e. the share of companies for which at least one establishment could be assigned, for the 571,662 companies in the Orbis data that have a value for EMPL that is at least 1.[24] Table 1 shows that the matching success rate for companies with a value for EMPL larger than 5 is above **80%**.[25] This is particularly remarkable given the fact that due to the imposed requirement of uniqueness of the identifier variable combination in the Orbis data (see the discussion in section 5), close to 10% of all companies (i.e. those with frequent company names such as “Fischer GmbH”) could not even enter the linkage.[26]

Table 1 also indicates that the matching success rate is much lower for companies with zero employment or missing employment information (subsumed as “missing” in the Orbis data, around 44%), which is not a surprise since the share of companies without regular employment will be large in this group. We also see that the matching success rate is about 55% for companies with a value of EMPL between 1 and 5, a group that may still include many small family companies without any regular employees, and that it increases dramatically (to 80%) once we cross the threshold of about 5 employees. The table also shows that the matching success rate does not continue to increase much more when we

[22] The variable EMPL is defined as “national company employment”, i.e. total company employment in Germany.

[23] Except for companies that were founded after the latest available IAB data of June 30, 2013.

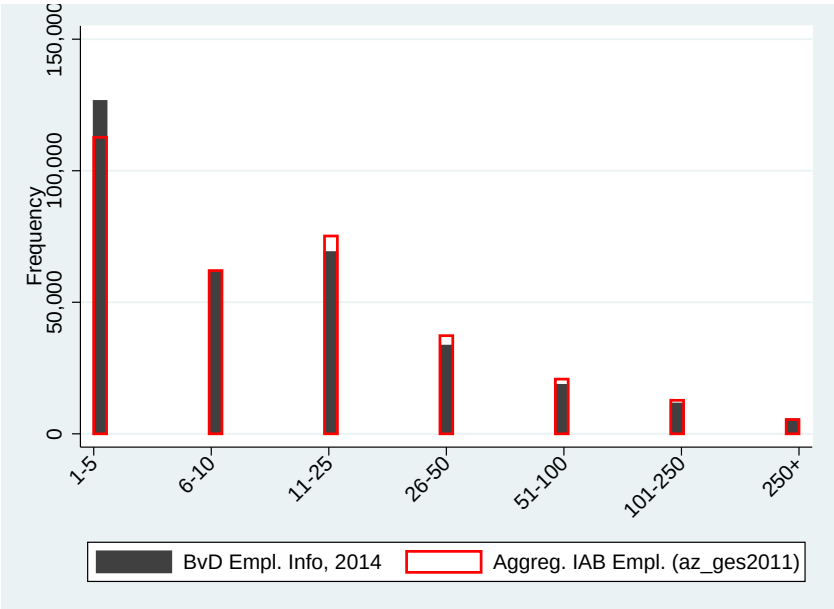
[24] Note that the Orbis variable EMPL is either “missing” (which includes 0) or larger than 0.

[25] The overall matching success rate for the subset of Orbis companies with a value of EMPL larger than five (rows 2 to 7 in table 1) is 82.4%, as can be calculated from table 1.

[26] Again, note that this is not problematic for representativeness of the final matched sample as long as having a common name is not systematically related to other company properties.

go further up to the largest size classes (up to 87.5% for companies with more than 250 employees).[27]

Figure 2: Company Employment according to Orbis and according to company-level aggregation of IAB variable “az_ges”



We use the EMPL variable to make further assessments of the quality of the matching procedure. We do this by aggregating for our m:1 company to establishment assignment the IAB Establishment data variable “az_ges”[28] to the level of the Orbis company ID (BvDID). This generates another variable of company level employment. We expect this variable to differ from the EMPL values due to at least the following reasons: a) due to the differences in underlying measurement concepts discussed above[29], and b) EMPL was likely measured in 2013 or in one of the preceeding years (unknown), while the latest values for az_ges that were available for this record linkage project are from June 30, 2011. Figure 2 compares the distribution of both of these variables (in size categories). Even with the likely considerably different data generating processes and different points of measurement of both variables, the Orbis EMPL variable and the company level aggregated IAB establishment variable largely agree on the total number of companies that should be in each of the 7 size categories, which is a pleasing result.

To see how large the deviation of both measures is regarding which companies to put in which categories, we cross tabulate both size categories in table 2. Note that the deviations are remarkably small considering the measurement issues for EMPL and the different

[27] The latter aspect is particularly comforting since it is a clear indication that large companies, even though they typically have a large number of establishments, still seem only slightly more likely to be matched at least one establishment.

[28] The variable az_ges, “Anzahl der Beschäftigten insgesamt”, describes the total number of employees (employees liable to social security contributions) of the establishment.

[29] We do not know the sign of this bias, although we may suspect that the measurement concept for az_ges should lead to lower values than for EMPL.

Table 2: Company Employment according to Orbis (EMPL) and according to Company-level Aggregation of the IAB Employment Variable (az_ges)

	Size according to aggregated IAB variable "az_ges"							
	1-5	**6-10**	**11-25**	**26-50**	**51-100**	**101-250**	**250+**	**Total**
Size according to Orbis	%	%	%	%	%	%	%	%
1-5	**81.0**	31.1	13.3	8.7	6.6	5.1	4.1	38.5
6-10	15.4	**49.6**	15.5	3.2	1.7	1.3	0.8	18.9
11-25	2.8	18.0	**62.2**	18.8	4.1	1.8	1.2	21.3
26-50	0.5	1.0	8.1	**60.5**	16.4	2.3	1.0	10.3
51-100	0.2	0.2	0.7	7.9	**64.0**	11.7	1.4	5.7
101-250	0.1	0.1	0.2	0.7	6.6	**72.4**	8.2	3.5
250+	0.0	0.1	0.1	0.2	0.6	5.4	**83.3**	1.7
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Source: Own calculations based on Orbis company data and IAB establishment data.

underlying concepts and given the fact that there are a few years between both measurements. Overall, we interpret these unexpectedly small deviations as a strong sign of a very low rate of false assignments.

Table 3: Match Success for Companies Active 2014, with at least 5 Company Employees (EMPL), by Legal Form

	no (=0) or at least one (=1) establ. assigned					
Legal form of Firm	**0**		**1**		**Total**	
	%	N	%	N	%	N
AG	12.7	559	87.3	3,835	100.0	4,394
GmbH	17.7	40,559	82.3	189,199	100.0	229,758
GmbH u.Co.KG	19.9	7,746	80.1	31,108	100.0	38,854
KG	28.3	116	71.7	294	100.0	410
LTD u.Co.KG	15.0	18	85.0	102	100.0	120
Ltd.Company	15.4	19	84.6	104	100.0	123
OHG	25.2	62	74.8	184	100.0	246
UG	17.4	159	82.6	757	100.0	916
UG u.Co.KG	28.0	35	72.0	90	100.0	125
e.G.	24.5	614	75.5	1,895	100.0	2,509
e.K.	51.2	503	48.8	479	100.0	982
gGmbH	18.1	509	81.9	2,310	100.0	2,819
Total	18.1	50,899	81.9	230,357	100.0	281,256

Source: Own calculations based on Orbis company data and IAB establishment data.

We also assess whether there is systematic selection into the matched set of companies due to other characteristics. Table 3 provides an overview of the matching success rates by legal form for companies with 5 or more employees.[30] The table shows that for sole proprietorships (e.K.) it is much less likely to find at least one establishment (around 48%). This could be due to a large share of small family businesses within this group. For the other legal forms, there do not seem to be large differences regarding the matching success

[30] Note that the number of matched firms is slightly lower in table 3, which is due to the fact that some Orbis companies do not have legal form information.

(around 83%).

Finally, it may be instructive to look at a distribution of the number of establishments that is matched to companies. Table 5 (see section A) shows that the vast majority of companies in the Orbis data that were assigned at least one establishment, are assigned exactly one establishment, and that the number of companies with more than 4 establishments is in the order of 1%, which is in line with other empirical evidence.

7 Summary

We have, for the first time, linked company data from a public data source (Orbis database) with establishment data from the German Federal Employment Agency. To do this, we have applied a set of matching procedures that we have described and theoretically motivated. We have assessed the quality of our matching procedures using a subset of companies for which establishments can reasonably be expected to exist in the IAB data (Orbis companies with more than 5 employees), and we have shown that for around 82.4% of these companies, at least one IAB establishment could be assigned. Our assessment of the matching quality shows that a) above the size of 5 employees, matching success does not dramatically increase further with size (up to 87.5% for the largest company size class), b) total company employment based on Orbis employment information can be very closely replicated by aggregating the employment variable az_ges from the IAB establishment data to the company level, and that c) the matching success rate is not strongly correlated with the legal form of a company, except for (as to be expected) sole proprietorships.

The new dataset opens various new possibilities for analyses. The establishment history panel (BHP) of the IAB data add accurate and yearly, longitudinal employment information that is not available in the Orbis data, including information on qualification, age, and other variables. With this new dataset, the financial variables in the Orbis data can be jointly analyzed with detailed longitudinal data on employment and occupational structure. This enables researchers, for example, to describe company productivity much more accurately, and to describe and to analyze joint changes in employment and financial variables on the company level over time.

Antoni et al. (forthcoming) describe the steps taken to create a research dataset from the key developed through this record linkage.

A Further Tables

Table 4: Matches by Match Method for Firms Active 2014

Matching Method	N%	%
(1) exact long name + legal form	506,870	31.1
(2) exact short name + legal form	117,102	7.2
(3) exact long name + 4dig.plz	64,128	3.9
(4) exact short name + 4dig.plz	23,643	1.5
(5) n-gram names (array), exact legal form	16,951	1.0
(6) exact long name + legal form + 1dig.industry	7,421	0.5
(7) exact short name + legal form + 1dig.industry	5,154	0.3
(8) exact long name + place	1,952	0.1
(9) exact short name + place	1,045	0.1
(10) rare name component triplets + 4dig.plz	72,576	4.5
(11) rare name component triplets + 3dig.plz	1,167	0.1
(12) rare name component triplets + place	3,692	0.2
(13) n-gram names (array), exact 4dig.plz	6,699	0.4
(14) exact long name w/o activity comp., 4dig.plz	12,767	0.8
(15) exact short name w/o activity comp., 4dig.plz	3,376	0.2
(16) exact long name w/o activity comp., 3dig.plz	1,635	0.1
(17) exact name w/o activity comp., 3dig.plz	362	0.0
(18) only in BvD Data	754,066	46.3
(19) reclass. to 'no match' by subsample regr.	27,062	1.7
Total	1,627,668	100.0

Source: Own calculations based on Orbis company data and IAB establishment data.

Table 5: Matches by No. of Assigned Establ. for Firms Active 2014

Nr. of Assigned Establ.	%	N
0	48.0	781,134
1	43.9	713,086
2	5.8	93,950
3	1.2	20,079
4	0.4	6,987
5	0.2	3,489
6-10	0.2	3,446
10-100	0.2	3,280
100-1000	0.0	221
1000+	0.0	13
Total	100.0	1,625,685

Source: Own calculations based on Orbis company data and IAB establishment data.

References

Antoni, Manfred, Katharina Koller, Marie-Christine Laible, and Florian Zimmermann (forthcoming). *Orbis-ADIAB: From Record Linkage Key to Research Dataset*. FDZ-Methodenreport xx/2016 (en).

Gruhl, Anja, Alexandra Schmucker, and Stefan Seth (2012). *Das Betriebs-Historik-Panel 1975-2010*. FDZ-Datenreport 04/2012.

Herzog, Thomas N., Fritz J. Scheuren, and William E. Winkler (2007). *Data Quality and Record Linkage Techniques*. New York: Springer.

IHK Köln (2012). *Wie finde ich die richtige Firmierung für mein Unternehmen?* `http://www.ihk-koeln.de/upload/Voraussetzung_Firmierung_8901.pdf`. Industrie- und Handelskammer (IHK) Köln.

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

Schnell, R., T. Bachteler, and S. Bender (2003). "Record Linkage Using Error-prone Strings". In: *Proceedings of the Section on Survey Research Methods*, pp. 3713–3717.

Schnell Rainer; Bachteler, Tobias; Bender Stefan (2004). "A Toolbox for Record Linkage". In: *Austrian Journal of Statistics* 33.1/2, pp. 125–133.

Imprint

Publisher
German Record-Linkage Center
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Stefan Bender
Rainer Schnell

Technical production
Heiner Frank

Template layout
Christine Weidmann

Download
`http://fdz.iab.de`

ISBN
...

IMPRINT

Publisher

German Record-Linkage Center
Regensburger Str. 100
D-90478 Nuremberg

Editors

Rainer Schnell, Manfred Antoni

Template layout

Christine Weidmann

Download

www.record-linkage.de

The German Record Linkage Center was funded by the German Research Foundation (DFG).