DEUTSCHE
BUNDESBANK
EUROSYSTEM

Linking Deutsche Bundesbank Company Data

Technical Report 2023-05

Version: v2023-1-6

Deutsche Bundesbank, Research Data and Service Centre

Eniko Gábor-Tóth
Christopher-Johannes Schild
Susanne Walter

Research Data and
Service Centre

Abstract

We present a method of automatically linking several data sets on companies based on supervised machine learning. We employ this method to perform a record linkage of several company datasets used for research and analytical purposes at the Deutsche Bundesbank. The record linkage process involves comprehensive data pre-processing, blocking / indexing, construction of comparison features, training and testing of a supervised match classification model as well as post-processing to produce a company identifier mapping table for all internal and public company identifiers found in the data. The evaluation of our linkage method shows that the process yields precise match predictions with a sufficiently high coverage / recall to make full automation of company data linkage feasible for typical use cases in research and analytics.[1)]

Keywords: Company data, Record Linkage, Data Matching, Supervised Machine Learning

Version: v2023-1-6

Citation: Gábor-Tóth, E., Schild, C., and Walter, S. (2023). Linking Deutsche Bundesbank Company Data. Technical Report 2023-05. Deutsche Bundesbank, Research Data and Service Centre.

[1] We thank Hendrik Doll, Simone Schultz and Franco Wieser for substantial contributions to early versions of our record linkage process. We thank Stefan Bender, Martin Eisele, Dominik Elgg and Daniel Werner for valuable comments and suggestions. We thank Sandra Gottschalk and Georg Licht for kindly providing the company master data from the MUP (”Mannheimer Unternehmenspanel”).

Contents

1 Release notes

New in version v2023-1-6:

1. Removed column separator issues in input data (RIAD, BvD) that have previously led to data loss / incomplete coverage of entities in the corresponding input data (in the order of ~0.1%).
2. Fixed incorrectly formatted IDs in the the input data (most notably RIAD: check digit in Nehmernummer) that have previously led to incorrect ground truth as well as false negatives (by ID assignment) in the record linkage result (in the order of up to ~1% of all RIAD entries).
3. Fixed a georeferencing error that led to false assignments by name / city as well as a flawed ground truth (in the order of ~1%). See section "Data Cleaning".
4. Changed treatment of entries on establishments in BvD-Data: in the previous version, entries referring to establishments were kept in the data with their full establishment-ID (i.e. BvD-ID + 4-digit establishment suffix), in the current version, entries referring to establishments are kept in the data only with their BvD-ID (i.e. without the 4-digit establishment suffix). See section "Data Cleaning".
5. Changed string comparison algorithms fed to first and second level model. See section "Comparison Features".
6. Integrated more recent data from the sources AWMuS (MiDi, SITS / SIFCT), BvD, LEI, JANIS, RIAD / AnaCredit, URS. See section "Input Data".
7. Added data source "Schufa". See section "Input Data".

2 Introduction

Linking different company data can increase the potential of this data to support evidence-based policy making and research. The need for linked company data has strongly increased in recent years, especially following the financial crisis and more recently the COVID-19 crisis. In response to this increased need for more integrated data, the RDSC has started to build up a record linkage infrastructure.[2)]

The general record linkage problem for records on entities such as persons or companies is a well studied problem (Christen, 2012): When data sources do not have common unique identifying keys, alternative identifying variables, such as names, addresses, legal form or economic sector information can be used to identify different records, i.e. different representations of the same real-world entity. One of the main challenges of record linkage stems from the fact that the alternative identifying variables on entities often differ between datasets. For example, databases may focus their quality checks on different positions or quality assurance rules may simply be different. In most cases, it is not possible to fully standardize these variables. Records therefore often have to be compared using computationally costly string similarity measures such as string distance metrics, and rules have to be found to combine the different similarity measures to decide which record pairs likely constitute a match.

Our record linkage process involves comprehensive data pre-processing, blocking / indexing, construction of comparison features, training and testing of a supervised match classification model as well as post-processing steps. This report provides a detailed description and evaluation of the record linkage techniques used to link company datasets of the Deutsche Bundesbank.

The final result of our record linkage process are company pairs that were identified by the record linkage process as being a match. These company pairs are organized into so-called ID-linkage tables (our data product "IDLINK"), which are two-column tables that provide a correspondence between entities as identified by two different identifiers (IDs) (Gabor-Toth, Schild, and Walter, 2023a).[3)] The overlaps based on these ID-linkage tables are analyzed in detail in a separate technical report (Gabor-Toth, Schild, and Walter, 2023b), which complements this technical report.

The ID-linkage tables enable internal and external researchers and internal analysts to link RDSC company analytical datasets to each other and to external company datasets.[4)] While the technical report Gabor-Toth et al. (2023b) is aimed at analysts and researchers primarily interested in the overlaps that can be generated between the linked data by using our ID-linkage tables, this technical report is most useful for readers who are interested in the methodological aspects of our record linkage.[5)]

The generated ID-linkage tables enable further data integration efforts, such as consolidation of different sources and data enrichment. Researchers and analysts can use these tables to combine data on companies in new ways, ask new research questions or examine them from a different angle (consider for example combining data on firms' foreign subsidiaries with firm-level balance sheet data).[6)]

[2] An early version of this record linkage, based on a smaller set of datasets, has been presented by (C.-J. Schild and Schultz, 2016) and described in (C.-J. Schild, Schultz, and Wieser, 2017) and (Doll, Gabor-Toth, and Schild, 2021). Currently there are efforts in place to establish more widely used unique common company identifiers that will help the future identification of entities across multiple datasets, such as the "LEI" of the Global LEI Foundation or the ECB RIAD-ID. Within the Deutsche Bundesbank, RIAD-BBk aims to integrate most company master data.

[3] The record linkage method described in this technical report corresponds to IDLINK-version v2023-1-6.

[4] For external researchers, these ID-linkage tables are anonymized in cases where one or both IDs are public or quasi public IDs.

[5] It may also be interesting for readers who would like to gain a better understanding of practical applications of supervised machine learning.

[6] Furthermore, our record linkage allows duplicate detection within data sources. Within Deutsche Bundesbank, we provide lists of duplicate candidates upon request.

Table 1: Analytical Datasets Linked

Dataset	Name	Period	Description	Master dataset
AnaCredit	Analytical Credit Data	2019-2021	Loan by loan data on credits larger than 25000 EUR.	RIAD
JANIS	Individual financial statements of non-financial firms	1997-2019	Annual financial statements of German non-financial companies. Successor to USTAN.	JANIS
MiDi	Microdatabase Direct Investment	1999-2020	Foreign direct investment (FDI) stock relations.	AWMuS
BAKIS-M	Millionenkreditevidenz	2002-2018	Borrower-lender level data on credit relationships of 1 million EUR or more.	BAKIS-M
SIFCT	Statistics on international financial and capital transactions	2001-2021	Microdata for the compilation of the financial account, capital account and investment income of the German balance of payments statistics.	AWMuS
SITS	Statistics on international trade in services	2001-2021	Microdata on international trade in services collected by the Deutsche Bundesbank.	AWMuS
USTAN	Corporate balance sheets	1987-2018	Annual financial statements of German non-financial companies. Predecessor to JANIS.	CoPS / JALYS
SCHUFA	Schufa Reference Data on Borrowers	2015-2022	Financial statement data of non-financial companies (only individual accounts) from Schufa.	SCHUFA

Note:
Time periods reflect the time intervals for which master data information was used for the current version of the record linkage processes. This might differ from the time coverage of the most recent versions of the analytical datasets. This is attributable to the fact that the record linkage processes are not restarted on every occasion when a new update for an analytical dataset is released. The end point of the time interval corresponds to the latest year for which observations were available in the standardized version of a particular dataset. The starting year shown for each dataset corresponds to the first year when observations for at least 10% of the average number of unique IDs are present in that year. The record linkage process, however, takes into account all available observations. For example, the Schufa data cover data from 2009 onwards but the number of IDs does not exceed 10% in these years. The entities in the years before 2015 are still considered in the record linkage.

3 Input Data

Deutsche Bundesbank company data are generated from a diverse set of company data sources, some originally collected for statistical reporting, others for prudential purposes. For the current record linkage application we augment Bundesbank datasets with other external company data. These come from Bureau Van Dijk (BvD, a commercial data provider), the SCHUFA Holding AG (a commercial data provider), public sources (LEI-data) and from the official business register (URS) of the German Federal Statistical Office (DESTATIS). We distinguish between analytical datasets, research datasets and master datasets. Our analytical datasets are snapshots of statistical and analytical databases that store information reported to the Deutsche Bundesbank to generate statistical aggregates or for prudential purposes. "Research datasets" are likewise snapshots of statistical and analytical databases, but different from (purely) analytical datasets, they additionally have to be anonymized, documented and versioned. Therefore "research data" may be seen as a subcategory of "analytical data".

Each analytical dataset (from here on meant as "including research datasets")[7] can be used in conjunction with exactly one master dataset, however multiple analytical datasets may be associated with the same master dataset[8]. The ID that links an analytical dataset and a master dataset is referred to as the dataset's "native ID".

Table 1, 'Analytical Datasets Linked', gives an overview on the analytical data.[9] The analytical datasets are described in detail in their corresponding dataset documentation or in research articles about these

[7] For convenience, for the rest of this technical report, we use the term "analytical data" as an upper category which includes "research data".
[8] Corresponding to an n:1 relationship.
[9] A more comprehensive overview is provided in the technical report Gabor-Toth et al. (2023b).

data.[10]. Each analytical dataset is linked to a master dataset, which enables identification of the entities in the analytical data for purposes of this record linkage.

Table 2: Master Datasets Linked

Dataset	Name	Period	Description	ID
AWMuS	Foreign Trade Statistics Reference Data	1981-2023	Repository for all foreign trade statistics related master and metadata in the Deutsche Bundesbank. Source of master data for MiDi, SITS and SIFCT.	MLD_NR
BAKIS-M	Bank Supervision Reference Data on Borrowers	2002-2018	Repository with master data on all borrower entities with a large credit satisfying the reporting requirements to the Deutsche Bundesbank as defined in the KWG. Apart from the borrower-lender level master data it also contains information on their credit of 1 Million or more. Source of master data for the research dataset generated from BAKIS-M.	DE_BAKISN_CD
BvD	Bureau Van Dijk Reference Data	2004-2023	Dataset with master data on non-financial companies, acquired from the external data provider ”Bureau Van Dijk”, complemented by the master dataset ”Mannheimer Unternehmenspanel” (MUP), from the Zentrum für Europäische Wirtschaftsforschung (ZEW).	BVD_CD
CoPS / JALYS / USTAN (earlier database)	CoCAS Providing System	1980-2018	Repository with HGB and IFRS annual financial statements for companies, insolvency data, data reported for the credit register and rating information, earlier in the context of refinancing operations and later for credit assessment purposes. Apart from this financial data, it contains master data on companies that have been reported to the Deutsche Bundesbank in this context. Prior to 1998, balance sheet data and the accompanying master data on companies was collected by a database also called ”USTAN” (not to be confused with the research dataset ”USTAN” that still exists, and that was named after this database). From 1998 on, balance sheet data as well as accompanying master data collection was transferred from USTAN to JALYS (later to be replaced by the database ”CoPS”). The database CoPS and their predecessors are the source of master data for the research dataset USTAN.	USTAN_CD
JANIS	Individual financial statements of non-financial firms	1997-2021	Annual financial statements of German non-financial corporations. Successor to USTAN.	USTANPLUS_CD
LEI	LEI Reference Data	2018-2023	Dataset with company master data by the Global Legal Entity Identifier Foundation (GLEIF).	LEI
RIAD	Register for Institutions and Affiliates Data	2019-2023	Central repository with master data for various Organisational Units and their relationships. Typically it contains more information about financial entities than non-financial entities. Source of master data for AnaCredit.	ENTTY_RIAD_CD
SCHUFA	Schufa Reference Data on Borrowers	2015-2022	Repository for all companies that are registered in the German Trade registry. It covers the master data for 1.9 Mio active and registered enterprises.	SCHUFA_ID
URS	Business register	2012-2021	Contains master data corresponding to the business register of the Federal Statistical Office of Germany.	WE_ID_ALT

Note:
Time periods reflect the time intervals for which master data information was used for the current version of the record linkage processes. This might differ from the time coverage of the most recent versions of the analytical datasets. This is attributable to the fact that the record linkage processes are not restarted on every occasion when a new update for an analytical dataset is released.The end point of the time interval corresponds to the latest year for which observations were avaialble in the standardized version of a particular dataset. The starting year for each dataset corresponds to the first year when observations for at least 10% of the average number of unique IDs are present in that year.

Table 2, ‘Master Datasets Linked’, gives an overview on the master data that enters the record linkage. These master databases are necessary to link the analytical data since analytical datasets do not contain

[10] For AnaCredit: Krodel, Orben, and Schild (2023), JANIS: Becker, Biewen, Schultz, and Weissbecker (2023), MiDi: Blank, Lipponer, Schild, and Scholz (2020), BAKIS-M: Schmieder (2006), SIFCT: Biewen, Pham-Dao, and Stahl (2022), SITS: Biewen and Lohner (2022), USTAN: Becker, Biewen, Schultz, and Weissbecker (2020)

identifiying attributes such as identifiers, names and addresses. For most of the master databases that we rely on, there also exists some form of written documentation that is publicly available.[11]

The usefulness of each identifier and each alternative identifying variable to link the different company datasets depends not only on its quality, but also on the degree to which it is available within each dataset as well as across datasets. For example, if only a small percentage of records in a dataset has a trade register ID, or if only a few datasets record the trade register ID, the potential of this attribute to aid in matching the data is limited. Table 3, 'Filling Ratios for (selected) Identifying Positions in the Input Master Data, by Dataset', shows, for the most important IDs and alternative identifying positions, to which degree these positions are available in the various datasets.[12]

Looking at Table 3, we make the following observations:

1. Most master datasets "main" ID (or "native ID") is only available in its native master data. Exceptions: a notable share of records in USTAN and in RIAD contain the BAKISN_CD, a notable share of JANIS records contain the USTAN_CD, and some records in USTAN, RIAD and JANIS contain the LEI code. These ID references are due to previous attempts to integrate parts of these datasets.[13]
2. The trade register ID occurs to a notable extent in every dataset, however, its prevalence is far from complete.
3. The prevalence of the trade register court information is a somewhat lower than that of the trade register ID. This is likely due to erronoeus or outdated textual information on trade register courts.
4. The prevalence of company name and address information is very high, altough there are some exceptions.
5. Email and telephone numbers are only present in a few datasets, with Email address prevalence being very low.
6. Economic Sector information is usually present, although filling ratios show a large variation.
7. Other alternative identifiers, such as founding years and economic sectors, are sufficiently prevalent to also contribute somewhat to the matching quality, especially when name or address information may be of limited quality.

Given the almost complete prevalence for names and addresses and given the limited overlap of common IDs, there is potential for finding additional matches using alternative identifiers. How large this potential is, in the end, depends on data quality and the true overlaps between the datasets, about which we have only incomplete knowledge: this knowledge is mostly theoretical and mostly consists of pre-conceptions of the specific data universes, given, for example, reporting requirements for specific data collections. Data overlaps and linkage success, with reference to the record linkage process described in this paper, are further discussed in Gabor-Toth et al. (2023b).

[11] BAKIS: Wehlert and Ißbrücker (2020), BvD: https://www.bvdinfo.com/en-gb/our-products/data, and, for Creditreform-Data, which is a large source for BvD-data that for Germany basically uses the same identifier: Bersch, Gottschalk, Müller, and Niefert (2014), CoPS/JALYS: *Benutzerhandbuch Für JALYS (WEB) Der Deutschen Bundesbank* (2007) and *Benutzerhandbuch CoPS (CoCAS Providing System)* (2020), JANIS: Becker, Biewen, Schultz, and Weissbecker (2019), LEI: https://www.gleif.org/en/about-lei/code-lists/gleif-registration-authorities-list/ and https://www.gleif.org/en/lei-data/lei-mapping/, RIAD: ECB RIAD Team (2019), URS: DESTATIS (2019).

[12] These alternative identifying positions are explained in Table 5 'Identifying Positions' in the appendix.

[13] They are of course very valuable for the record linkage process, and they are used, for example, to generate training and testing data for our classification model.

Table 3: Filling Ratios for (selected) Identifying Positions in the Input Master Data, by Dataset

	BVD	BAKIS-M	AWMuS	LEI	URS	RIAD	USTAN	JANIS	SCHUFA
BVD_CD	1		0.05					0.78	
DE_BAKISN_CD		1				0.47	0.22		
AWMUS_CD			1						
LEI	0.03			1		0.14	0.03	0.02	
DE_DESTATIS_CD_STBL					1				
ENTTY_RIAD_CD						1			
USTAN_CD							1	0.27	
JANIS_CD								1	
SCHUFA_ID									1
REG_ID	0.64	0.25	0.52	1	0.66	0.94	0.07	0.61	1
REG_LCTN_DE	0.59	0.08	0.51	0.01	0.66	0.94	0.07	0.61	0.98
NM_ENTTY_ASC_CL	1	1	1	1	1	1	1	1	1
LGL_FRM	0.64	0.64	0.62	0.73	0.77	0.93	0.65	0.02	0.89
PSTL_CD	0.97	0.37	0.12	1	1	1	0.41	0.99	1
CTY_NORM	0.93	1	1	0.99	1	1	0.99	1	1
STRT_NORM	0.77		0.06	0.95	0.96	0.88	0.11	0.34	0.97
HSNR	0.75		0.06	0.93	0.99	0.85	0.1	0.33	0.95
EML	0.1		0.03						
TLFN	0.18		0.4				0.01		
DT_BRTH_YR	0.7				1	1	0.51	0.97	
ECNMC_ACTVTY	0.85		0.84		1	0.92	0.16	1	1
Total	7060320	1461551	492778	187860	3343362	1335598	266463	484273	1724686

Note:

The meaning of the positions can be found in Table 5, 'Identifying Positions', in the appendix.

The universes comprise observation deduplicated over time and only german entities.

Due to random missingness in the input data, some observations might be still duplicated over the years.

4 Data Cleaning

Each dataset is standardized before entering the record linkage. This affects all the variable names and for some variables, their content. Variable names are standardized according to the data standard defined by AnaCredit, currently v2.2.[14]. Standardization at the value level covers standardizing value meanings (codelists) for categorical variables and defining similar scales and units for continuous variables. If possible, standardization of values occurs also according to the data standard defined by AnaCredit, currently v2.2.[15].

Names

Firm names in the Bundesbank databases originate from either paper or electronic forms submitted to the Bundesbank. Their quality depends on a number of different factors, such as the frequency and quality of manual or automatic cross-checks with other data sources. Errors such as typos in company names are present in the Bundesbank datasets as in all external datasets used for the record linkage. Typical issues with company names are non-harmonized abbreviations, uninformative insertions of name components of different kinds, as well as typing errors (single letter insertions, deletions etc.). For the firm name fields, data cleaning involves removing known variation in different correct notations, such as standardizing the German word "Gesellschaft" to its most common abbreviation "Ges" and "&", "+", "und", "and" etc to "UND". It also involves replacing German Umlauts "ä", "ö", "ü" by their common non-Umlaut replacements "ae", "oe", "ue" as well as capitalizing. Legal form information is extracted from the firm name field and removed from the firm name (see below).

[14] https://www.bundesbank.de/de/service/meldewesen/bankenstatistik/formate-xml/formate-xml--611846
[15] https://www.bundesbank.de/de/service/meldewesen/bankenstatistik/formate-xml/formate-xml--611846

Legal Form

German company names should contain the legal form. Depending on the source, the full legal form is spelled out or it is abbreviated in some way. To detect the many different ways to abbreviate legal form information, a set of regular expressions was developed, repeatedly tested and improved until more than 95% of the legal form information was detected correctly. Most databases also included a coded variable for the legal form. The codelists for this original legal form information differ strongly with regard to granularity and are therefore harmonized to rather coarse categories.

Addresses

Addresses are validated and standardized to their official spelling according to the address reference dataset for Germany made available by the Federal Agency for Cartography and Geodesy ("Bundesamt für Kartographie und Geodäsie"), using the software "infas 360 PAGS Geocoder"[16]. In some specific cases, the georeferencing software comes up with incorrect municipality codes ('AGS': Amtlicher Gemeindeschlüssel). These errors occur in rare cases when both a) the address contains only the city name (no postal code, street) and b) this city name is either not unique (such as "Neustadt") or spelled ambiguously (such as "Frankfurt" instead of "Frankfurt am Main" or "Frankfurt (Oder)"). In theses cases all entries were previously falsely assigned to the municipal code of the first city name in the list (according to the correct spelling, in this above example, "Frankfurt (Oder)"). For this record linkage, this issue was fixed by first performing an empirical check of the frequencies of city name / AGS-combinations and then assigning city names without postal code information an AGS based on theses frequencies, i.e. only if a city name spelling is almost exclusively linked to one AGS (in at least 98% of all cases).

IDs

External firm IDs present in more than one dataset were standardized if they followed different conventions in the different databases (such as the case for trade register court identification). Outdated trade register information, for example due to moved headquarters or changed legal forms, was corrected using trade register legacy tables, which were derived from reports of trade register ID changes ("Handelsregistermeldungen"). Likewise, outdated BvD-ID values were corrected using a BvD-ID legacy table. The BvD dataset contains also information on establishments. In the previous version of our record linkage process, BvD-IDs for establishments appeared in their original form, BvD-ID plus the 4-digit establishment suffix. This has in many cases led to multiple assigments based on probabilistic matching between various establishments and the corresponding parent company (which would only appear with its BvD-ID, if such an ID is present in the dataset). Such multiple assignments are deleted during post-processing and thus the previous handling of establishment level data has led to a loss in our record linkage results. In the current version however, we remove the 4-digit establishment suffixes in the cleaned input data for the record linkage. This ensures that the entries refer to the legal entity behind the establishment, not the establishment itself. This way the number of multiple assigments is reduced and coverage of our current record linkage version is increased.

[16] www.infas360.de

Other Variables

Information on founding dates were coarsed to the founding year, telephone number formats were harmonized.

5 Blocking

To reduce the number of match candidate pairs while trying to preserve as many record pairs as possible that refer to the same entity, we apply "indexing" or "blocking". This consists in applying inexpensive exact pre-filters in order to sufficiently reduce the number of costly comparisons while at the same time blocking out as few true matches as possible (Christen, 2012).

A complete classifier compares every representation found in the datasets with every other representation since it is not known ex ante, which pairs of representations can be ruled out to be a match. For s datasets with n representations each, this would however result in a very large number of comparisons. Even with moderately large data, it becomes quickly unfeasible to run computationally costly comparisons on all these pairs. In order to apply many costly comparisons, such as string distance measures, one therefore has to limit the search space for matching pairs. Since there are no exact filters available that can be expected to be entirely free of errors (even the postal code may be erroneous and subject to change), several blocking rules are (additively) combined, accepting pairs that match on either of these rules. The filter variables we consider as blocking criteria are generated from 1. cleaned company name, 2. cleaned company name tokens, truncated[17], 3. city, 4. postal code, 5. street name, 6. NACE Rev. 2 sector code (2 digits), 7. telephone, 8. founding year, 9. legal form.

Most blocking keys are generated by combining name-based block components (1) or (2) with one or several of the other, not name-based blocking components.[18] Variation is introduced through a combination of components as well as different lengths of prefixes for the single block components. For example, the cleaned company name's first 5 letters are combined with the first 2 digits of the postal code. Blocking keys may also be entirely non-name-based or entirely name based, such as, for example, concatenating the founding year, the postal code and the first digit of the sector code or by combining the first and the second name component. This procedure generates a total of 1,130 blocking keys, that are quite diverse with regard to their strictness, i.e. the positions they are based on, the prefixes used from each position as well as the number of comparisons they generate. Of these, a total of 200 rules is randomly[19] selected.[20]

Before joining on each selected blocking key, block sizes are calculated for each key. If the blocksize for any block within any key[21] is larger than a pre-defined maximum block size value x[22], i.e. if more than x entries are in the block, the blocking key is not used. In such cases, we have to trust that entries within this block will likely be found by a different, stricter blocking key definition[23].

[17] Company name components truncated to 6 characters.
[18] Blocking key components are concatenated as strings to form a blocking key.
[19] One exception being that the first 35 characters of the cleaned firm name are always selected as a blocking key.
[20] We have experimented with changing the number of randomly selected blocking rules and found that increases in the number of blocked candidates does not increase much more after about 100 selected rules.
[21] For example, firm names starting with common name component prefixes as "SCHMID" in combination with the city name "BERLIN"
[22] For this record linkage, a maximum block size of 50 was chosen.
[23] Such as, for example, firm names starting with "SCHMID" in combination with the city name "BERLIN" in combination with a second, more specific name component or the economic sector code, the founding year etc.

Overall, the blocking procedure reduces the number of comparisons from the order of roughly $N = 10^{13}$ to about $C = 10^8$ candidate pairs. This would lead to a reduction ratio of $RR = 1C/N = 99.999\%$, i.e. for all following steps, we limit our search for matching pairs to 0.001% of all theoretically possible matching pairs.

6 Comparison Features

In order to inform our classification model, to enable it to learn which similarity patterns between two match candidates are typically associated with a match, and which are typically associated with a non-match, respectively, we need a set of meaningful comparison features that capture different aspects of similarity between two given company records. The similarity measures are mostly based on company names and addresses, but also on a few other attributes. The comparison features most relevant for our record linkage are listed and described in Table 4, 'Construction of Comparison Features', on page 13. ##### Name-based Clearly one of the most important positions to distinguish the different firm entities is the firm name field. Different variants of the preprocessed company names were compared: the original, non-standardized firm name, the standardized firm name without legal form information, the standardized firm name up to the position where the detected legal form information begins, as well as a concatenatation of the single tokens of the firm name, with each token being truncated to the first 6 characters.

As string comparison metrics[24], we use the "fuzzy ratio" from the FuzzyWuzzy Package for python (Seatgeek, 2020), the Levensthein distance and the "Generalized Edit Distance" as implemented in SAS by the functions "COMPLEV" and "COMPGED"[25] (Lafler and Sloan, 2018), Jaro and Jaro-Winkler metrics from the Jellyfish package for Python (Turk, 2020) as well as bigram and trigram measures, from the Natural Language Processing toolkit for Python (Bird, Klein, and Loper, 2009). All features based on string comparisons used in this record linkage are listed in Table 6 on page 24 in the appendix. The set of features differs slightly from the set of features from the previous version. The reason is that the set of string comparison algorithms fed to first and second level models has been changed changed.[26]

Within company groups, sometimes companies are distinguished merely by adding a suffix corresponding for example to a within group running numerical id to the firm name (e.g. 1,2,3). For string comparison algorithms, this often leads to a misleadingly high similarity of the string, since often only one (numeric) character differs in these cases, but it makes all the difference to distinguish these companies. To use a fictional example, there may be the entries "1. ABC Real Estate Investment GmbH" and "2. ABC Real Estate Investment GmbH", as well as name versions with roman numbers or even spelled out numbers, such as "Erste ABC Real Estate Investment GmbH", "ABC Real Estate Investment II GmbH" and so on. To give our prediciton model a chance to distinguish such entries, and to attribute meaning to enumerations within firm names, we extract numbers (arabic, roman or spelled out) from the firm names and generate a comparison feature that explictly indicates for each match candidate pair whether there are numbers found in both firm names and if so, whether these numbers are matching or non-matching.

[24] For a general overview on string comparison metrics see Cohen, Ravikumar, and Fienberg (2003) and Christen (2012).

[25] COMPGED is a metric that punishes for insertions, deletions, replacements differently based on a cost function that was derived from best practice experiences in general data matching tasks (Lafler and Sloan, 2018).

[26] Most importantly, comparatively slow cosine metrics from the machine learning library sci-kit learn (from which we use both word and character-based metrics) were dropped and replaced by additional metrics based on the FuzzyWuzzy Package for python, due to the availability of a c-compiled (thus faster) version of the fuzzy package.

Table 4: Construction of Comparison Features

Feature input position	Feature input category	Feature input description	Comparisons for 1st level model	Comparisons for 2nd level model
AGS	Address	Municipality Code	exact, jaro-winkler	
CTY_NORM	Address	City (standardized)	exact, edit distance	
HSNR	Address	Housenumber	exact	
INFID	Address	Address-ID	exact	
PSTL_CD	Address	Postal Code	exact, jaro-winkler	
STRT_NORM	Address	Street (standardized)	exact, edit distance	
WGS84_Xdez, WGS84_Ydez	Address	Coordinates	exact, euclidian distance	
ECNMC_ACTVTY2_consKUSY	Economic Sector	Bundesbank Codelist	exact	
ECNMC_ACTVTY2_consNACE	Economic Sector	Nace v2 Codelist	exact	
EML	Email	Email	exact	
DT_BRTH_YR	Founding Date	Founding Year	exact	
LGL_FRM	Legal Form	Legal Form according to AnaCredit Codelist	exact	
LGL_FRM_RDSC	Legal Form	Legal Form according to RDSC Codelist	exact	
blockcomp1	Name	First name component	exact, levenshtein	
blockcomp2	Name	Second name compontent	exact, levenshtein	
blockcomp3	Name	Third name component	exact, levenshtein	
FDSZ_NumbersInName	Name	First digit extracted from the firm name	exact	
NM_ENTTY_ASC_NL_CL	Name	Cleaned name w/o legal form	exact, levenshtein	fuzzy ratio, fuzzy partial ratio, fuzzy token sort ratio, fuzzy token set ratio
NM_ENTTY_ASC_UL_CL	Name	Cleaned name until legal form	exact, edit distance	
NM_ENTTY_ASCrfextr	Legal Form	Legal form extracted from firm name	exact	
NM_ENTTY_BLCK_CNCAT	Name	Truncated name components, concatenated	exact, edit distance, jaro	
NM_ENTTY_BLCK_CNCAT_SPACE	Name	Truncated name components, concatenated w spaces	exact, bigrams, trigrams, fuzzy ratio, fuzzy partial ratio, fuzzy token sort ratio, fuzzy token set ratio	
NM_ENTTYcl	Name	Cleaned name, including spelled legal form	exact, jaro, edit distance	
TLFN	Telephone	Telephone number	exact, jaro-winkler	
probRF, probXGB, probLR	Predictions from 1st level models	Scores from models using all 1st level features		model score

Location-based

For German resident companies, especially for smaller firms, the company name alone cannot be expected to be unique within Germany, therefore information on the location of the firm is needed to uniquely identifiy the firms (Schäffler, 2014). Georeferenced locations were compared by exact comparison of the full standardized address-ID. To account for the possibility of renamed streets or changes of firm location within cities, postal codes and municipality codes were also compared. Because of often false house number information or formatting, and to circumvent cases of failed georeferencing, we also included string distances on the city name and on the street name. To account for the possibility of location changes across municipality borders, we also included euclidian distance between geographical coordinates.

Other Features

As additional features, we compare codified legal forms, legal form information extracted from the firm names using regular expressions (see section "data cleaning"), exact comparison of the founding year, as well as the first and second digit of the sector code.

7 Groundtruth

To train a classifier, we need a subset of matching pairs for which it is known whether these pairs really constitute a match or not ("groundtruth"). In order to allow the classifier to discover as many general relations between feature value similarities and the match status of a match candidate pair as possible, this subset has to be as large and as representative of the universe of potential matches as possible. To test the prediction quality of a classifier, another subset of such pre-existing knowledge is needed. This subset has to be kept separately and should not be used to train the classifier up to the point of validation. While certainty about the true match status of any given match candidate pair is hardly achievable, we can at least derive a sort of "quasi groundtruth". This quasi groundtruth can then be used for model training on the condition that we calibrate our model so that it is insensitive to outliers.

We generate the groundtruth dataset based on common IDs. The rules to derive candidate pairs for our groundtruth are the same as described below in the section "Match classification", subsection "ID-based Match Classification". The number of match candidate pairs in the original groundtruth that are defined by our ID-based match classification rules as true negatives is considerably larger than the number of true positives. In order to have more balanced training and test data, we therefore undersample true negatives for our final groundtruth used for training and testing. The overall size of the final groundtruth dataset used for training and testing is 3, 207, 067[27], the share of true positives in this final groundtruth amounts to 37.3%.

Training / Test Split

Based on the concatenated fields "cleaned firm name" and "municipality code", we use a hash algorithm to generically tag roughly 15.6% of the match candidate pairs in the groundtruth as "hold-out" pairs, i.e. pairs[28] that are not to be used in training at all, and only to be used for final testing.

Since our modelling involves a stacked model (see below), of the remaining groundtruth, 75% are first used for training several first level (or "base"-) models. The remaining 25% of the groundtruth serves

[27] This number is only about 60% of the value published in the previous version of this technical report, TR 2021-05, which referred to the record linkage version v2021-2-6. This is likely due to several factors (compare the "release notes" to this version): 1. previously probabilistic matches where the city only differed due to the (now fixed) false georeferencing (cases with different city but otherwise very similar master data) that are now exact matches 2. changed treatment of entries on establishments in BvD-Data: since in the previous version, entries referring to establishments were kept in the data with their full establishment-ID (i.e. BvD-ID + 4-digit establishment suffix), instead of, as in the current version, with their BvD-ID (i.e. without the 4-digit establishment suffix), this leads to an overall considerably smaller contribution of the BvD-Data to the groundtruth.

[28] As defined by their cleaned firm names and municipality codes.

a dual purpose: first it is used to test the first level models, then it is used for training the second level stacked model.

8 Match Prediction

For all match candidate pairs, we predict a match probability using a supervised classification model. For all modelling calculations, the python package "Scikit-learn" is used (Pedregosa et al., 2011).

8.1 Model

The classification model is a "stacked" model which consists of two model layers or "levels": several first level models and a second level model. There are two reasons why we chose this design. First, since we have more than 120 million match candidates, computational costs for calculating comparison features for all pairs are quite large. This is especially true for computationally costly comparison features, like some string comparison algorithms. A stacked model enables us to sort out a lot of very unlikely matches by using predictions from one (best performing) first level model as a filter, reducing the number of pairs that enter the second model to about one third of the original number of match candidates. For these pairs, we then calculate the more expensive string algorithms, which, together with the first level model scores, then enter the second model as features. Second, since we have plenty of training data, stacking models can be expected to improve model performance, since they are often able to capture interactions between the first level base models, which often have particular strengths and weaknesses in capturing certain patterns in the data (Breiman, 1996; Wolpert, 1992).

First Level Model

The first level "base"-models are: 1. random forest (Breiman, 2001) 2. "extreme gradient boosting trees"-model (XGBoost) (Chen and Guestrin, 2016) 3. logistic regression.

Hyperparameter search is done using a randomized parameter grid search algorithm, as proposed by (Bergstra and Bengio, 2012) and implemented in Pedregosa et al. (2011), within a set search space,[29] and standard values as implemented in "Scikit-learn" otherwise (Pedregosa et al., 2011), using 5-fold cross-validation to find the best parameters. To guarantee balanced splits in the cross-validation samples, and to further reduce calculation time, we chose the stratified shuffled split cross validation estimator.

Since we want to find a large share of correct matches, but also many matches, we are both interested in a large precision and a large recall (see below in the section "Evaluation"). This means that we want our model to be balanced in the sense that it neither focusses too much on precision at the cost of a low recall nor vice versa. Therefore, we choose the target score to maximize in the parameter search to be the F1-Score, which is the harmonic mean of precision and recall. $F1 = 2 \cdot (precision \cdot recall)/(precision + recall)$

[29] For the random forest, maximum depth is set between 10 and 25, the number of estimators is fixed to 200, the minimum samples in a split and the minimum samples in a leaf are between 2 and 9. For the XGB model, maximum depth is set between 10 and 25, the number of estimators is between 50 and 100, the learning rate is between 0.1 an 0.2, and the subsample restriction is between 0.5 and 0.7. For the logistic regression, C is between 1 and 10, with a 'sag' solver. Features are set to the same scale for the logistic regression using the robust scaler algorithm.

Figure 1, 'Feature Importance for 1st Level Random Forest Model', on page 17 shows feature importance scores for the random forest base model fed with all available features[30]. Feature importance on the x-axis is measured as Gini-Importance (as described by Breiman (2001)), so values correspond to shares. We observe that location and name-based features clearly are the most important group of features by this measure. This was to be expected given the presumption that firm names and location are the most important alternative features commonly used to distinguish different companies. We also observe that feature importance seems to be strongly correlated with the filling ratios of the positions from which the comparison features are generated[31], which was also to be expected.

Second Level Model

The second level takes the first level model scores as features, plus 4 string comparison features from the 'fuzzy' package, applied to a different name variant 'NM_ENTTY_ASC_NL_CL', i.e. the company name stripped of the legal form, but including name components that occur after the legal form.[32].

For each of the three first level base models, we calculate two variants, one using all available features[33], and then one model variant that takes a randomly limited subset of these features. This gives us 6 first level model scores to use as features in the second level, plus 3 string comparison features not used in the first level models (see above). The second level comparison features and scores are only calculated for pairs that score above a certain threshold in the (best performing) first level model. This threshold is chosen such that about one third of the pairs make it to the second level.[34]

Figure 2 shows feature importance scores for the features used in the second level model. In Figure 2, as in Figure 1, feature importance on the x-axis is measured as Gini-Importance (as described by Breiman (2001)). While Figure 2 seems to show that the name features that were only included for the second level add little to model performance on their own, we still keep them in the model, since leaving them out leads to a slightly lower overall perfomance of our final second level model.

8.2 Evaluation

We evaluate our predictor model using a holdout set (see section "Groundtruth") of the data on known match / non-match pairs that were not fed to the classifier for training and which do not constitute exact matches. This means that this hold-out set constitutes "unseen data" to the model and to the entire human - machine interactive process of calibrating the model up to this point. We compute predictions for the pairs in the hold-out set and compare these predictions with their true match / non-match status. This leads to four possible outcomes:

- The pair is a non-match and correctly classified as a non-match ("true negative", TN)

[30] Construction of comparison features from the available alternative identifying variables is described in Table 4, a list of all comparison features can be found in Table 6, 'Features Based on Continuous Comparison Metrics', on page 24 and Table 7, 'Features Based on Exact Agreement', on page 25 in the appendix.
[31] See Table 3, 'Filling Ratios for (selected) Identifying Positions in the Input Master Data, by Dataset', on page 9
[32] These comparison features are described in Table 6, 'Features Based on Continuous Comparison Metrics', on page 24 in the appendix.
[33] i.e. all features are described in Tables 7 and 6 in the appendix.
[34] Originally, up until the previous version of this record linkage, the rationale for this was to only calculate these (previously) expensive string comparisions only for those pairs that score above a certain threshold in the (best performing) first level model. Since this version, we use a faster (c-compiled) version of the fuzzy package. For convencience, and due to the fact that this two-level "stacked model"-architecture still performs better than a one-level architecture, we stick to the stacked model.

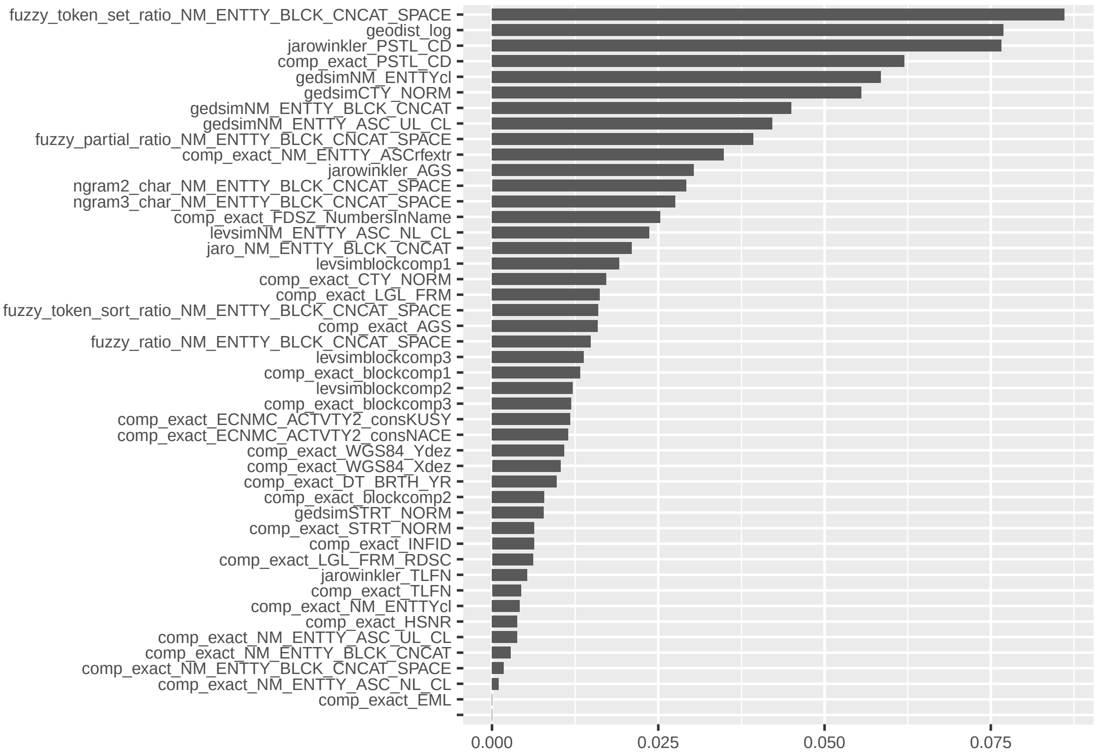

Figure 1: Feature Importance for 1st Level Random Forest Model

Note: Feature acronyms with the prefix 'comp_exact' are based on exact comparisons of the corresponding feature input variables and are described in Table 7, 'Features Based on Exact Agreement', on page 25 in the appendix. All other features are based on continuous comparison metrics (mostly string comparison metrics) and are described in Table 6, 'Features Based on Continuous Comparison Metrics', on page 24 in the appendix.

– The pair is a non-match and incorrectly classified as a match ("false positive", FP)
– The pair is a match and correctly classified as a match ("true positive", TP)
– The pair is a match and incorrectly classified as a non-match ("false negative", FN).

Two measures are used for evaluation: Precision is defined as the fraction of true positives over all pairs classified as matches by the classifier, i.e.: TP/(TP + FP). Put differently, it is the share of the classified matches (TP + FP), that are in fact matches (TP). Recall is defined as the fraction of true positives over all known true matches, i.e. TP/(TP + FN), or: the share of the known true matches (TP + FN) that the classifier classified correctly (TP). Since each classified pair is assigned a matching likelihood by the classifier, we can trade precision against recall by changing the likelihood threshold above which a pair is classified as a match. Depending on our relative preferences for precision and recall, which depends on the analytical question, it may either be more desirable to include a rather large share of true matches in the analysis, at the expense of a correspondingly large share of false positives (high recall / low precision), or it may be more desirable to include a rather low share of false positives, at the expense of missing a relatively large number of true matches. We finally chose a threshold of 0.75, putting the emphasis a bit more on "precision" than on "recall".

The set of achievable combinations of precision and recall is described by Figure 3 (for the second level model (orange) and the first level random forest model (blue)).

The second level model seemingly shows a better performance when compared to the first level model. This is to be expected from model stacking, however, it is important to note that the two precision recall

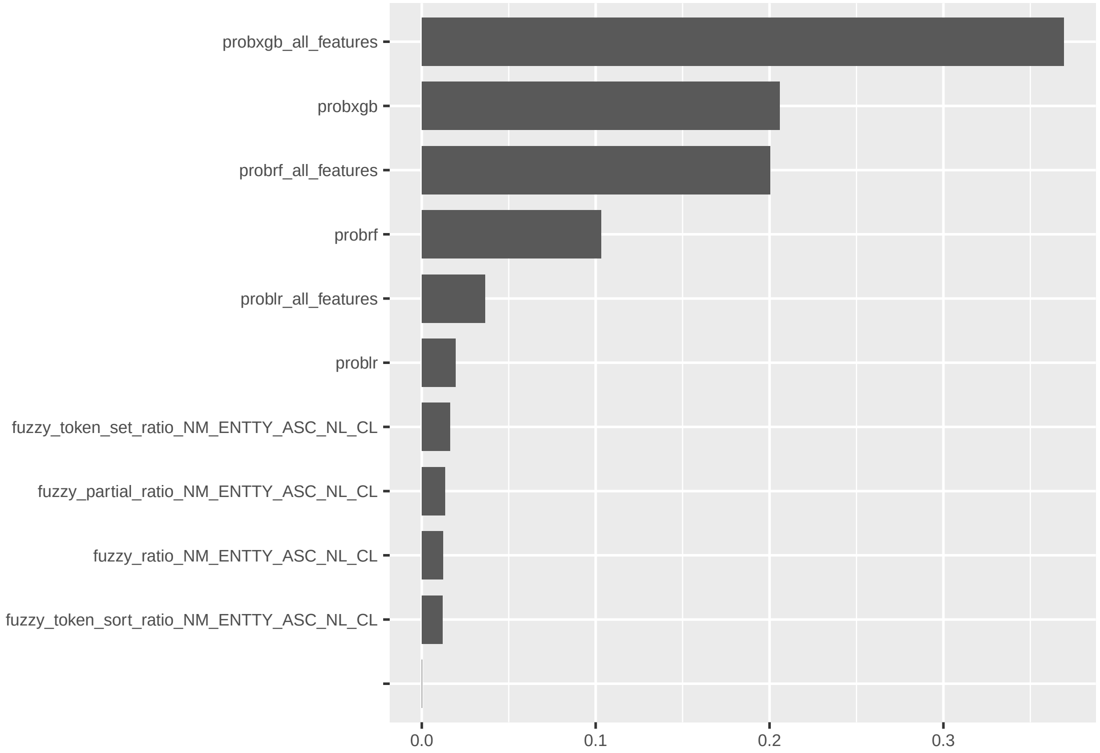

Figure 2: Feature Importance for the 2nd Level Random Forest Model

Note: 'probxgb_all_features': score from random forest model using all 1st level features, 'probrf_all_features': score from random forest model using all 1st level features, 'probrf': score from xgb model using a random subset of the 1st level features, 'probxgb': score from xgb model using a random subset of the 1st level features, 'problr_all_features': score from logistic regression model using all 1st level features, 'problr': score from logistic regression model using a random feature subset of the 1st level features. For 'cosine_char_wbNM_ENTTY_BLCK_CNCAT', 'fuzzy_ratio_NM_ENTTY_ASC_UL_CL' and 'cosine_tokenNM_ENTTY_BLCK_CNCAT' see Table 6, 'Features Based on Continuous Comparison Metrics', on page 24 in the appendix.

curves presented here are not directly comparable since the second level model, due to the elimination of pairs with a low first level score, uses a limited subsample of the groundtruth to evaluate model performance (i.e. only pairs that score high in the first level models).

9 Match Consolidation

In order to make a final decision for each match candidate pair on whether it should be classified as a match or not, we do not only use the model prediction, but we make use of all available information derived from model prediction, common IDs and exact name and address matches:

1) Exact agreement on name and address
2) Common Identifiers
3) Match prediction model score

Exact Matches

Match candidates that are exactly identical based on both the cleaned firm name and the municipality code found by address georeferencing, are classified as exact matches. Exact matches do not enter model

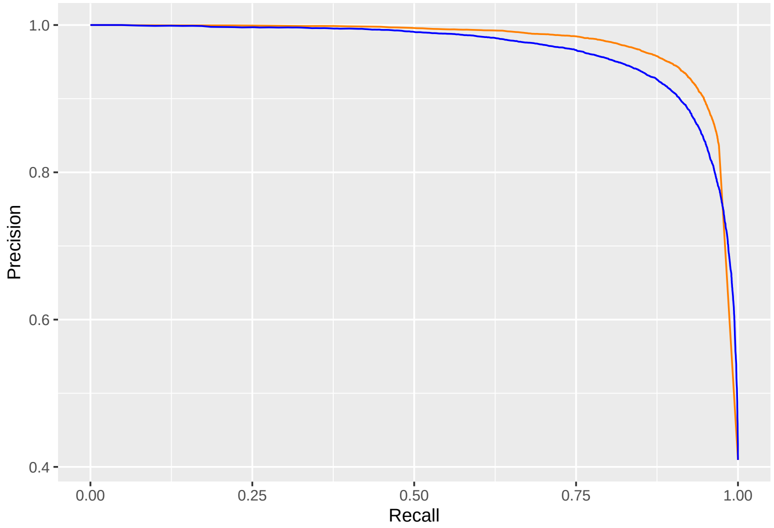

Figure 3: Precision / Recall Curves, 1st (blue) and 2nd Level (orange) Model

scoring. If the exact name / place rule signals a match, but the the ID-based matching rules do not, then the exact name / place rule overrides the ID-based match classification.

ID-Based Match Classification

Generally, match candidate pairs with a common identical internal or external ID, and without contradicting match information with regard to any other internal or external ID, are classified as matches. There are however exceptions to this rule. A matching "DE_BAKISN_CD", "USTAN_CD", "DE_TAX_CD" or "DE_VAT_CD" only count as an ID match if at least one other other common ID also matches (it may also be one of the just mentioned IDs). This accounts for substantial differences of the entity concept behind the respective IDs[35] and known (rare) cases of re-assigned historic ID values[36]. Also, there are IDs which we trust to signal a match even if there is contradicting match information from other IDs.[37]

Match candidate pairs are also classified as final non-matches based on common ID information. Due to ID formatting heterogeneities and issues of outdated ID information, we are more hesitant to classify a pair as a non-match based on non-matching common IDs. Generally, we limit the set of common IDs trusted to provide any signal for non-matches[38]. The general rule is as follows: if no common ID matches, and if at least two IDs trusted to signal non-match information are filled and have different values, we classfiy the pairs as a non-match based on ID information.[39] There is however an exception also to this rule: there are IDs which we trust to signal a non-match on their own, i.e. even if there is no other non-matching ID confirming this non-match.[40]

[35] This is the case for "DE_BAKISN_CD", "DE_TAX_CD" and "DE_VAT_CD".
[36] This is the case for "USTAN_CD".
[37] This is the case for "LEI", "BVD_CD", "AWMUS_CD", "REG_ID_LCTN" and "ENTTY_RIAD_CD".
[38] These are the IDs "LEI", "BVD_CD", "DUNS_CD", "AWMUS_CD", "DE_BAKISN_CD" and "REG_ID" (without the alphanumeric ending)
[39] Note that, due to the priority of the exact match rule over all ID-based rules, such a pair may still be classified as a match based on the cleaned name and the municipality code, see above.
[40] This is the case for "LEI" and "BVD_CD".

Score-Based Match Classification

The third set of rules is based on the probabilistic score from the machine learning model described in the previous section. It is applied only when the first two rules yield no result. Whenever neither the exact name/place rule nor the ID-based rules come to the conclusion of either a match or a non-match, then all pairs with a score above the threshold of 0.9 are classified as matches. There are, however, exceptions to this rule: whenever for at least one of the two match candidates, there are other match candidates that match[41], then such a probabilistic match is discarded. As a result, probabilistic matches are only accepted for mutually best match candidates. This means that in the final ID-linkage tables, there exist no multiple ID assignments that are based on probabilistic matching.

Final Evaluation of the Match Quality

Since the final match classification is affected by our post-processing rules, and since the evaluation based on test data is limited by the assumption that the available test data is representative, we want to take a closer look at the quality of our final match results. To do this, we 1) briefly investigate the effects of our post-processing rules on the final match quality and 2) manually evaluate a random subsample from our final match classification table.

Final match classification is based on consolidating all derived match indicators from ID comparisons, exact correspondence of name and address, and the probabilistic model scores. First of all, it is important to note that due to the fact that, at least in absolute terms, there is a substantial amount of information on especially external IDs in the data, it is not surprising that final classification in the end rests to a large extent on ID comparisons. Another large share of matches can be assigned by exact name and address comparisons.

Looking at the most relevant relations among all bilateral relations between the matched candidates,[42], we find that the total number of matches found for these data source relations amounts to $17,725,385$. For these most relevant relations, we find that on average, 17.1% of all matches are based on IDs, 80.4% of all matches are exact matches based on cleaned names and addresses, and the remaining 2.6% are probabilistic matches. Although seemingly low[43], this result is in fact good news for the overall match quality, since it seems reasonable to assume that ID-based matches and exact matches can be considered more reliable than probabilistic matches. At the same time, the high share of exact matches based on names and addresses is suggestive evidence for a successful data pre-processing and standardization.

Furthermore, the additional complexity that probabilistic matching brings along is compensated and justified by manifold analytical reasons. Firstly, probabilistic matching is in the first place necessary for measuring the share of probabilistic assignments and a reduction of its share over several years. Secondly,

[41] Either according to another probabilistic score above the threshold or according to the exact name/place rule or the ID-based rules

[42] Here defined as all relations within the set of internal datasets, plus all relations between every internal data and the largest external datasources URS, BvD and Schufa, plus the relation between these external datasources. This corresponds to the following datasets: BAKIS, AWMUS, URS, BVD, USTAN, JANIS, RIAD, SCHUFA.

[43] This share has noticeably decreased since the previous version, where 5.1% were probabilistic matches. This decrease may partly due to the fact that in the current record linkage, we have excluded more than 2 million BvD-Establishments, which more often than other records do not match exactly (or by ID) to their legal entity. To a small extent, it is also due to the new datasource "Schufa" not included previously: without Schufa, the share of probabilistic assignments is slightly larger (2.9%). Additionally, it may be due to previously incorrectly formatted IDs which are now more often ID-matches instead of previously probabilistic matches. Note that a mismatch in the Nehmernummer (which was strongly affected) alone does not result in an ID-nonmatch if there is no mismatch in a another common ID: therefore it may well result in an exact or a probabilistic match.

given that it can be important for research and analysis to get as close as possible to a complete matching, even an additional 2.6 percentage points may be valuable. Thirdly, the share of probabilistic matches varies across bilateral relations: its standard deviation is 1.8%-points. Its maximum value is 5.92% (for the relation USTAN - URS), which in absolute terms corresponds to 7, 868 matches.[44]

Finally, in absolute terms, for this record linkage the number of probabilistic matches amounts to 291, 665.[45] So these matches would practically not have been found without the probabilistic component of our record linkage.

To evaluate overall match quality, we manually review a random subsample from our final match classification table. Out of 1000 randomly chosen classified pairs, fewer than 10 are false positives, which leaves us with an overall precision of more than 99%. This overall precision value is larger than the precision calculated for our matching model, which is due to the fact that the share of probabilisitc matches in the final matching results is relatively small, as described above.

10 Conclusion

Records in analytical datasets and in the corresponding company master datasets used at Deutsche Bundesbank cannot be linked easily through unique common IDs. To enable researchers and analysts to use linked company data, the RDSC regularly matches company data using current record linkage techniques. In the present report we have described our record linkage processes for company data. This report helps users gain a better understanding of the complexitity of this record linkage process, the approach taken by the RDSC, and it also facilitates a better interpretation and usage of our results. For a thorough description and interpretation of the result of our record linkage, the ID-linktables data product "IDLINK", we refer the reader to the accompanying technical report (Gabor-Toth et al., 2023b).

Contact

Deutsche Bundesbank, Research Data and Service Centre (RDSC)
E-mail: fdsz-data@bundesbank.de
Homepage: https://www.bundesbank.de/rdsc

Citation requirements

For any study or other document which is made available to the public and contains information derived from the provided data, the researcher is obliged to properly cite the data source as:

Gábor-Tóth, E., Schild, C., and Walter, S. (2023). Linking Deutsche Bundesbank Company Data. Technical Report 2023-05. Deutsche Bundesbank, Research Data and Service Centre.

[44] Its minimum value is 0.04% (for the relation SCHUFA - BVD), which in absolute terms still corresponds to 834 matches. Factors likely to drive up the probabilistic share could be scarce availability of common identifiers, low quality of names and addresses, a large share of old / outdated information in the data and also, quite generally, independent of quality, lack of bilateral similarity between databases, for example due to relatively little (substantial) past bilateral data exchanges between the databases involved.

[45] We estimate that for a person working full time on manually evaluating match candidates in order to find additional matches (i.e. not matchable by ID or exact comparison) it would take around 2, 917 working days to go through the list of potential match candidates.

References

Becker, T., Biewen, E., Schultz, S., and Weissbecker, M. (2019). *Individual Financial Statements of Non-financial Firms (JANIS) 1997-2017* [Data Report 2019-10].

Becker, T., Biewen, E., Schultz, S., and Weissbecker, M. (2020). *Corporate Balance Sheets (Ustan) 1987-2019* [Data Report 2020-08].

Becker, T., Biewen, E., Schultz, S., and Weissbecker, M. (2023). *Individual Financial Statements of Non-financial Firms (JANIS) 1997-2022* [Data Report 203-14].

Benutzerhandbuch CoPS (CoCAS Providing System). (2020). (5-514).

Benutzerhandbuch für JALYS (WEB) der Deutschen Bundesbank. (2007). (1.3).

Bergstra, J., and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*(10), 281–305.

Bersch, J., Gottschalk, S., Müller, B., and Niefert, M. (2014). *The Mannheim Enterprise Panel (MUP) and Firm Statistics for Germany* [Discussion Paper]. (14-104).

Biewen, E., and Lohner, S. (2022). *Statistics on International Trade in Services (SITS) 01/2001 - 12/2021* [Data Report 2022-17].

Biewen, E., Pham-Dao, L., and Stahl, H. (2022). *Statistics on International Financial and Capital Transactions (SIFCT)* [Data Report 2022-07].

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with python*. O'Reilly.

Blank, S., Lipponer, A., Schild, C.-J., and Scholz, D. (2020). Microdatabase Direct Investment (MiDi) – A Full Survey of German Inward and Outward Investment. *German Economic Review*, (1).

Breiman, L. (1996). Stacked Regressions. *Machine Learning*, *24*(1), 49–64.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In B. Krishnapuram, M. Shah, A. J. Smola, C. Aggarwal, D. Shen, and R. Rastogi (Eds.), *KDD* (pp. 785–794). ACM.

Christen, P. (2012). *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection* (1st ed.). Berlin Heidelberg: Springer.

Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003). A Comparison of String Distance Metrics for Name-Matching Tasks. *Proceedings of the 2003 International Conference on Information Integration on the Web*, 73–78. Acapulco, Mexico: AAAI Press.

DESTATIS. (2019). *Unternehmensregistersystem - Qualitätsbericht*.

Doll, H., Gabor-Toth, E., and Schild, C.-J. (2021). *Linking Deutsche Bundesbank Company Data* [Technical Report 2021-05].

ECB RIAD Team. (2019). *RIAD Reference Manual*. (RHN19.72-2).

Gabor-Toth, E., Schild, C.-J., and Walter, S. (2023a). *Company (ID) Linktables - (IDLINK)* [Data Report 2023-15].

Gabor-Toth, E., Schild, C.-J., and Walter, S. (2023b). *Understanding Overlaps between Different Company Data* [Technical Report 2023-06].

Krodel, T., Orben, J., and Schild, C.-J. (2023). *AnaCredit - German Part* [Data Report 2023-X]. (forthcoming).

Lafler, K. P., and Sloan, S. (2018). *A quick look at fuzzy matching programming techniques using SAS software* (No. MWSUG Paper SB-010). MidWest SAS Users Group.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Schäffler, J. (2014). *Re(LOC) Linkage: a New Method for Linking Firm-level Data with the Establishment-level Data of the IAB* [FDZ Methodenreport]. (201405_en).

Schild, C.-J., and Schultz, S. (2016). Linking Deutsche Bundesbank Company Data Using Machine-Learning-Based Classification: Extended Abstract. *Proceedings of the Second International Workshop on Data Science for Macro-Modeling*. New York, NY, USA: ACM. https://doi.org/10.1145/2951894.

2951896

Schild, C.-J., Schultz, S., and Wieser, F. (2017). *Linking Deutsche Bundesbank Company Data using Machine-Learning-Based Classification* [Technical Report 2017-02].

Schmieder, C. (2006). The Deutsche Bundesbank's Large Credit Database BAKIS-M and MiMiK. *Schmollers Jahrbuch*, (126).

Seatgeek. (2020). FuzzyWuzzy: Fuzzy String Matching in Python (Version 0.17.0). Retrieved from https://github.com/seatgeek/fuzzywuzzy

Turk, J. (2020). Jellyfish: a Library of Functions for Approximate and Phonetic Matching of Strings (Version 0.8.2). Retrieved from https://github.com/jamesturk/jellyfish

Wehlert, B., and Ißbrücker, M. (2020). *Meldetechnische Durchführungsbestimmung für die Abgabe der Großkreditanzeigen nach Art. 394 CRR (Stammdaten- und Einreichungsverfahren) und der Millionenkreditanzeigen nach §14 KWG*.

Wolpert, D. H. (1992). Stacked Generalization. *Neural Networks*, *5*, 241–259.

A Further Tables

Table 5: Identifying Positions

Variable	Short description
AWMUS_CD	AWMuS identifier. Anonymous. Original name: MLD_NR
BVD_CD	Bureau van Dijk identifier. Original name: bvdid
CTY	City
DE_BAKISN_CD	Borrower ID ("Nehmernummer")
DE_DESTATIS_CD_STBL	Destatis business register ID. Original name: WE_ID_ALT
DT_BRTH	Founding Year
ECNMC_ACTVTY	Economic Sector (Nace v2)
EML	Email
ENTTY_RIAD_CD	RIAD identifier.
HSNR	House Number
JANIS_CD	JANIS identifier. Original name: poolid
LEI	Legal entity identifier
LGL_FRM	Legal Form for Company, according to AnaCredit Technical Specifications, v2.2
NM_ENTTY	Name of Company
PSTL_CD	Postal Code
REG_ID	Trade register number, according to AnaCredit Technical Specifications, v2.2
REG_LCTN	Trade register court ID ("Justiz-ID"), according to AnaCredit Technical Specifications, v2.2
SCHUFA_ID	SCHUFA Holding AG identifier. Original name: SCHUFA_UNTERNEHMEN_ID
STRT	Street
TLFN	Telephone
USTAN_CD	USTAN identifier. Original name: ukn

Table 6: Features Based on Continuous Comparison Metrics

Comparison feature	Compared variable(s)	Comparison metric[a]
fuzzy_partial_ratio_NM_ENTTY_ASC_NL_CL	Cleaned firm name w/o legal form	Fuzzy Partial Ratio
fuzzy_partial_ratio_NM_ENTTY_ASC_UL_CL	Cleaned firm name until legal form	Fuzzy Partial Ratio
fuzzy_partial_ratio_NM_ENTTY_BLCK_CNCAT_SPACE	Truncated firm name components, with spaces	Fuzzy Partial Ratio
fuzzy_ratio_NM_ENTTY_ASC_NL_CL	Cleaned firm name w/o legal form	Fuzzy Ratio
fuzzy_ratio_NM_ENTTY_ASC_UL_CL	Cleaned firm name until legal form	Fuzzy Ratio
fuzzy_ratio_NM_ENTTY_BLCK_CNCAT_SPACE	Truncated firm name components, with spaces	Fuzzy Ratio
fuzzy_token_set_ratio_NM_ENTTY_ASC_NL_CL	Cleaned firm name w/o legal form	Fuzzy Token Set Ratio
fuzzy_token_set_ratio_NM_ENTTY_ASC_UL_CL	Cleaned firm name until legal form	Fuzzy Token Set Ratio
fuzzy_token_set_ratio_NM_ENTTY_BLCK_CNCAT_SPACE	Truncated firm name components, with spaces	Fuzzy Token Set Ratio
fuzzy_token_sort_ratio_NM_ENTTY_ASC_NL_CL	Cleaned firm name w/o legal form	Fuzzy Token Sort Ratio
fuzzy_token_sort_ratio_NM_ENTTY_ASC_UL_CL	Cleaned firm name until legal form	Fuzzy Token Sort Ratio
fuzzy_token_sort_ratio_NM_ENTTY_BLCK_CNCAT_SPACE	Truncated firm name components, with spaces	Fuzzy Token Sort Ratio
gedsimCTY_NORM	City, standardized	Generalized Edit Distance
gedsimNM_ENTTY_ASC_UL_CL	Cleaned firm name until legal form	Generalized Edit Distance
gedsimNM_ENTTY_BLCK_CNCAT	Truncated firm name components, concatenated	Generalized Edit Distance
gedsimNM_ENTTYcl	Cleaned firm name including spelled legal form	Generalized Edit Distance
gedsimSTRT_NORM	Street, standardized	Generalized Edit Distance
geodist_log	X-Coordinate (WGS84), Y-Coordinate (WGS84)	Log of euclidian distance
jaro_NM_ENTTY_BLCK_CNCAT	Truncated firm name components, concatenated	Jaro Distance
jaro_NM_ENTTYcl	Cleaned firm name, including spelled legal form	Jaro Distance
jarowinkler_AGS	Municipality Code	Jaro Winkler Distance
jarowinkler_PSTL_CD	Postal Code	Jaro Winkler Distance
jarowinkler_TLFN	Telephone number	Jaro Winkler Distance
levsim_NM_ENTTY_ASC_UL_CL	Cleaned firm name until legal form	Levenshtein Distance
levsimblockcomp1	First component of the firm name	Levenshtein Distance
levsimblockcomp2	Second component of the firm name	Levenshtein Distance
levsimblockcomp3	Third component of the firm name	Levenshtein Distance
levsimNM_ENTTY_ASC_NL_CL	Cleaned firm name w/o legal form	Levenshtein Distance
ngram2_char_NM_ENTTY_ASC_NL_CL	Cleaned firm name w/o legal form	Bigram Distance
ngram2_char_NM_ENTTY_BLCK_CNCAT	Truncated firm name components, concatenated	Bigram Distance
ngram2_char_NM_ENTTY_BLCK_CNCAT_SPACE	Truncated firm name components, with spaces	Bigram Distance
ngram3_char_NM_ENTTY_ASC_NL_CL	Cleaned firm name w/o legal form	Trigram Distance
ngram3_char_NM_ENTTY_BLCK_CNCAT	Truncated firm name components, concatenated	Trigram Distance
ngram3_char_NM_ENTTY_BLCK_CNCAT_SPACE	Truncated firm name components, with spaces	Trigram Distance

[a] Sources for comparison metrics are referenced in section 5.

Table 7: Features Based on Exact Agreement

Comparison feature	Compared variable(s)
comp_exact_AGS	Municipality Code
comp_exact_blockcomp1	First component of the firm name
comp_exact_blockcomp2	Second component of the firm name
comp_exact_blockcomp3	Third component of the firm name
comp_exact_CTY_NORM	City, standardized
comp_exact_DT_BRTH_YR	Founding Year
comp_exact_ECNMC_ACTVTY2_consKUSY	Nace v2, 2 digits, consolidated w. preference to internal sources
comp_exact_ECNMC_ACTVTY2_consNACE	Nace v2, 2 digits, consolidated w. preference to external sources
comp_exact_EML	Email address
comp_exact_FDSZ_NumbersInName	First digit extracted from the name
comp_exact_HSNR	Housenumber
comp_exact_INFID	Address-ID
comp_exact_LGL_FRM	Legal Form, RIAD Code
comp_exact_LGL_FRM_RDSC	Legal Form, RDSC Code
comp_exact_NM_ENTTY_ASC_NL_CL	Cleaned firm name w/o legal form
comp_exact_NM_ENTTY_ASC_UL_CL	Cleaned firm name until legal form
comp_exact_NM_ENTTY_ASCrfextr	Legal form extracted from firm name
comp_exact_NM_ENTTY_BLCK_CNCAT	Truncated firm name components, concatenated
comp_exact_NM_ENTTY_BLCK_CNCAT_SPACE	Truncated firm name components, with spaces
comp_exact_NM_ENTTYcl	Cleaned firm name incl. spelled legal form
comp_exact_PSTL_CD	Postal Code
comp_exact_STRT_NORM	Street, standardized
comp_exact_TLFN	Telephone number
comp_exact_WGS84_Xdez	X-Coordinates (WGS84)
comp_exact_WGS84_Ydez	Y-Coordinates (WGS84)

Note:
All comparison features are boolean variables (i.e. 1 ('identical values'), 0 ('different values') or 'missing').