DEUTSCHE BUNDESBANK
EUROSYSTEM

Understanding Overlaps between Different Company Data

Technical Report 2021-06

Version: v2021-2-6

Deutsche Bundesbank, Research Data and Service Centre

Eniko Gábor-Tóth
Christopher-Johannes Schild

Research Data and Service Centre

Abstract

We analyze overlaps between various company datasets, building on the results of the company data record linkage by Doll, Gabor-Toth, & Schild (2021). To better understand the data overlaps, we also briefly describe the input data for this linkage, in particular with respect to data universes and time periods covered by the data. We report descriptive statistics that characterize the overlaps found between the company data. The overlaps are discussed and interpreted with reference to properties of the input data and of the record linkage process.[1)]

Keywords: Company data, ID Linkage Tables, Record Linkage, Data Matching

Version: v2021-2-6

Citation: Eniko Gábor-Tóth, Christopher-Johannes Schild (2021), Understanding Overlaps between Different Company Data, Technical Report 2021-06 – Version v2021-2-6. Deutsche Bundesbank, Research Data and Service Centre.

1 We thank Stefan Bender, Martin Eisele, Dominik Elgg, Miro Holzer, Katharina Muno, Susanne Walter, Michael Weissbecker, Daniel Werner and Katja Ziprik for valuable comments and suggestions. We thank Hendrik Doll and Siddharth Marya for assistance with the multilateral overlap graphs, Michaela Krayer for her support with the SIFCT data and Sandra Gottschalk and Georg Licht for kindly providing the company master data from the MUP ("Mannheimer Unternehmenspanel"). We also thank Katja Ziprik and Katharina Muno for providing additional match information produced by the RIAD Team of Deutsche Bundesbank.

Contents

1 Introduction

Several departments of the Deutsche Bundesbank collect microdata on companies for different analytical and reporting purposes. From some of these databases, the Research Data and Service Centre (RDSC) of the Deutsche Bundesbank derives curated and documented datasets to be used for internal analysts and internal and external researchers. To enable data users to reap analytical value from linked company data, the RDSC matches them using current record linkage techniques.

Record linkage of company data is a complex problem. Due to the still insufficient use of common internal or public identifiers in the original master data, most of the analytical and research datasets cannot be linked easily through unique common IDs. To overcome this challenge we rely on comprehensive data manipulation steps and a supervised machine learning approach. The methodological details of our record linkage technique are described by Doll, Gabor-Toth, & Schild (2021).

The results of this record linkage are available for research and policy analysis as the data product "IDLINK," which corresponds to a collection of "ID-linkage tables." These ID-linkage tables are two-column tables containing a different company identifier in each column, resulting in a table of ID value pairs, that show which ID values of the two IDs refer to the same real-world company entity. The ID value pairs included in these linkage tables define overlaps between the different company datasets. This technical report describes and interprets these data overlaps. The list of ID-linkage tables is presented in more detail in Gabor-Toth & Schild (2021).

In Section 2, this report provides an overview of the datasets involved in the record linkage (Doll, Gabor-Toth, & Schild, 2021), which are consequently the also data basis for IDLINK (Gabor-Toth & Schild, 2021). In Section 3, we first present a short summary of the record linkage process and then move on to discuss and evaluate in detail the data overlaps generated by IDLINK. Section 4 concludes.

2 Company Data

2.1 Dataset Types

For the purpose of this technical report, 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 is meant as "including research datasets")[2] can be used in conjunction with exactly one master dataset, however multiple analytical datasets may be associated with the same master dataset[3]. The ID that links an analytical dataset and a master dataset is referred to as the dataset's "native ID."

The research datasets are described in detail in their corresponding dataset documentation or in research articles about these data. These documentations also provide a comprehensive description of the dataset specific universe.[4]. For other analytical datasets there is no corresponding

2 For convenience, for the rest of this technical report, we use the term "analytical data" as an upper category which includes "research data."
3 Corresponding to an n:1 relationship.
4 JANIS: Becker, Biewen, Schultz, & Weissbecker (2019a), MiDi: Blank, Lipponer, Schild, & Scholz (2020), SITS: Biewen & Lohner (2019), SICFT: Biewen & Stahl (2020)

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	1999-2018	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

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 research datasets. This is attributable to the fact that the record linkage processes are not restarted on every occasion when a new update for a research 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.

research dataset, this is for example the case for AnaCredit and BAKIS-M[5]. For an overview on all analytical datasets (including the research datasets), see Table 1, "Analytical Datasets Linked".

Master datasets are snapshots of master databases, i.e. internal databases that collect, integrate, process and provide identifying attributes and contact information on the company entities. These master databases are necessary to link the analytical data, since analytical datasets usually do not contain universal identifiying attributes such as universal 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.[6] For an overview on the scope of each dataset see Table 2, "Master Datasets Linked", on page 6.

2.2 Scope of the Datasets

2.3 Time Coverage

Table 2 on page 6 gives an overview on the time span for each dataset. The end point of the time interval corresponds to the latest year for which observations were available at the point when the master data snapshot was taken. 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.[7]

The length of the time spans covered differ between datasets, due to different scopes and the legal frameworks of the data collections. The time dimension is of relevance to evaluate the size of a particular dataset at a given point in time or when the goal is to characterize the overlap between certain datasets at again any point in time.

5 BAKIS-M: Wehlert & Ißbrücker (2020), AnaCredit: Alves-Werb et al. (2020)
6 BAKIS-M: Wehlert & Ißbrücker (2020), BvD: https://www.bvdinfo.com/en-gb/our-products/data - we complement BvD-data by data from the "Mannheimer Unternehmenspanel" (MUP) (Bersch, Gottschalk, Müller, & Niefert, 2014), CoPS/JALYS: "Benutzerhandbuch für JALYS (WEB) der Deutschen Bundesbank" (2007) and "Benutzerhandbuch CoPS (CoCAS Providing System)" (2020), JANIS: Becker, Biewen, Schultz, & Weissbecker (2019a), 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)
7 To calculate the dataset specific average number of unique IDs we take the number of unique IDs for each year averaged over the number of years covered by the observation period. This implies that for example for AWMuS observations are available already earlier than 1979 but 1979 is the first year when the number of unique IDs in that year is at least as much as 10% of the average number of unique IDs calculated over the full AWMuS time span.

Table 2: Master Datasets Linked

Dataset	Name	Period	Description	ID
AWMuS	Foreign Trade Statistics Reference Data	1980-2021	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	2004-2021	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	1997-2019	Annual financial statements of German non-financial corporations. Successor to USTAN.	USTANPLUS_CD
LEI	LEI Reference Data	2018-2018	Dataset with company master data by the Global Legal Entity Identifier Foundation (GLEIF).	LEI
RIAD	Register for Institutions and Affiliates Data	2019-2021	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
URS	Business register	2012-2019	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 research datasets. This is attributable to the fact that the record linkage processes are not restarted on every occasion when a new update for a research 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.

2.4 Universes

In the context of company data, a dataset universe is a set of real-world company entities that a given dataset is supposed to include information on. We do not aim to provide a full description of each dataset’s specific universe as this should be adressed in their respective documentation as cited in this document. However, since the information content and the scope of data collection affects which companies can be potentially matched across datasets, it is instructive to briefly review the type of companies that are contained in these datasets. The company universe of IDLINK covers all companies that appear at least once in any of the master datasets that have entered the record linkage for IDLINK.

SIFCT (AWMuS)

SIFTC contains information on international financial and capital transactions and serves the compilation of the financial account, capital account and investment income of the German balance of payments statistics. SIFCT is described in Biewen & Stahl (2020). The master data for this research dataset is maintained in AWMuS. AWMuS has been the common master database for all foreign statistics related master data of the Deutsche Bundesbank. It has originally been created (in the form of an initial load at the time of its creation in 2009) from its predecessor AUSUT and other previous foreign statistics databases. Therefore, even though AWMuS only exists since 2009, the earliest entries in AWMuS have a time reference that predates the year it was established.

MiDi (AWMuS)

MiDi is a dataset on foreign direct investment stock relations and is described in Blank, Lipponer, Schild, & Scholz (2020). Firms are required to report their cross-border shareholdings, covering both direct and indirect investments of non-residents in Germany, which allows the Bundesbank to generate corresponding statistics on international capital linkages.

SITS (AWMuS)

SITS is a dataset on international trade in services. It is based on information reported by German residents on service transactions above the statutory threshold which the Bundesbank uses to calculate the balance of payments statistics for Germany. The SITS is described in Biewen & Lohner (2019).

BAKIS-M (BAKIS-M)

BAKIS-M collects borrower-lender level data on individual credit relationships for so called large credits (as defined in Art. 394 CRR) and on individual credit relationships equal to or exceeding 1 Million Euro per borrower or borrower unit (as defined in § 14 KWG). BAKIS-M was set up to monitor the indebtedness of borrowers in Germany and the credit portfolio of banks. More details are provided in Schmieder (2006).

USTAN (COPS/JALYS)

USTAN is a dataset tat contains information based on firms' balance sheets and profit and loss accounts submitted directly to the Deutsche Bundesbank, initially for the credit-worthiness analysis performed at the Bundesbank in the context of its refinancing operations. Its universe is smaller than that of JANIS, however USTAN starts earlier than JANIS. The universe of USTAN is described in Becker, Biewen, Schultz, & Weissbecker (2019b). COPS (starting from April 2014) and JALYS (until April 2014) hold master data for the USTAN research dataset.

JANIS (JANIS)

The dataset JANIS contains balance sheet data for firms stemming from the credit-worthiness analysis performed at the Bundesbank in the context of its monetary policy operations. This main source of information is complemented with other internal and external sources. The dataset is documented by Becker, Biewen, Schultz, & Weissbecker (2019a). In the case of JANIS there is also a non-anonymized version of the dataset, that additionally also holds the master data for the

JANIS research dataset. This can be accessed from the premises of the Deutsche Bundesbank and is available for internal users only. JANIS is also the result of a comprehensive project aimed at further developing the USTAN dataset.

AnaCredit (RIAD)

AnaCredit is a dataset that contains detailed credit and credit risk information at borrower-by-borrower and loan-by-loan level for all loans above 25,000 Euro or more. The umbrella term "credit" is used as it collects information on multiple instruments, like loans, advances and bills of exchange. The universe of AnaCredit is described in detail in Alves-Werb et al. (2020). Master data for AnaCredit is collected and maintained in RIAD that supports and provides master data to several users within the Eurosystem, ESCB and SSM (for master datasets see Table 2). AnaCredit and RIAD are not yet available for general research purposes.

BvD

BvD[8] is a large master data set compiled and maintained by a commercial data provider. We complement BvD-data by master data from the "Mannheimer Unternehmenspanel" (MUP), which, since both databases use the ID of the data provider Creditreform, are linkable via their shared ID.[9] While the data from BvD available to the RDSC mostly refers to recent years, the MUP is a valuable complement, since it goes further back in time.

URS

URS[10] is the business register maintained at the German Federal Statistical Office, Destatis. It aims to comprise all companies in Germany that contribute to the German GDP in a particular year and is used for generating official statistics related to the economic activity in Germany. Further, URS is the most comprehensive master database on German companies with the German Federal Statistical Office.[11] More details are provided in DESTATIS (2019). URS and BvD contain the highest number of unique company IDs in any given year that is available for our record linkage purposes. URS enters our record linkage in order to allow the statistics department of Deutsche Bundesbank to use information from the URS, such as economic sector information, for analyses on data quality in Bundesbank company databases.

2.5 Total Number of Entites

Figure 1, "Unique entities by input dataset (pooled), total and filtered", on page 9 shows the number of companies in each dataset that has entered the record linkage that has generated IDLINK. This graph shows the size of both master datasets (AWMuS, BAKIS-M, BvD, JANIS, RIAD, URS) and selected[12] analytical datasets (MiDi, SITS, SIFCT, BAKIS-M, JANIS, USTAN). Because scope and time coverage differ, there is considerable variation in the size of these datasets, the smallest being MiDi, the largest being URS. Although MiDi has a longer time coverage, it only captures firms that engage in foreign direct investments or that are direct investment objects, whereas URS

8 Bureau van Dijk company master data - https://www.bvdinfo.com/en-gb/our-products/data.
9 Bersch, Gottschalk, Müller, & Niefert (2014)
10 "Unternehmensregister"
11 Not available for research. Since URS is used for statistical quality purposes at Deutsche Bundesbank, we nevertheless report case numbers and matching rates to URS.
12 URS is only available for statistical purposes but not for research purposes and information from BvD is only available for researchers who entered into a separate licence agreement with Bureau van Dijk, that allows the use of their data.

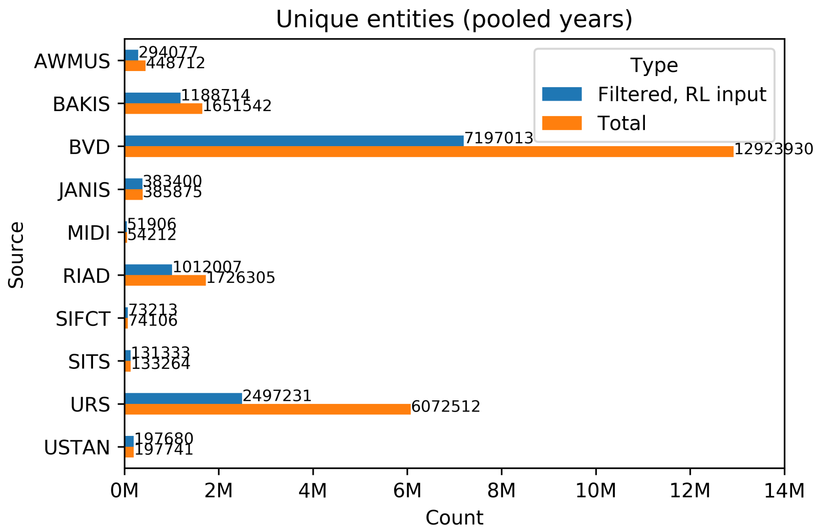

Figure 1: Unique entities by input dataset (pooled), total and filtered

aims at capturing all firms that contribute to the German GDP in a given year. Figure 1 illustrates that whereas information recorded in the master datasets AWMuS, BAKIS-M, JANIS and RIAD are tied to a more specific mandate of the Bundesbank, data collection for BvD and URS fulfill a more general purpose. This is reflected in the significant difference in their corresponding size.

Note that the dataset size illustrated with the orange bars corresponds to the datasets as they were originally transmitted to the RDSC. Since matching information on natural persons is not the focus of our record linkage, we remove observations that can be classified as natural persons.[13)] Additionally, to a smaller extent, observations are dropped if they do not contain a minimum set of identifying information[14)] or if the companies represented are not resident in Germany.[15)] This applied filters alter the size of the datasets as shown by the blue bars. The filtered datasets are the starting point for the record linkage.

The potential to classify observations as natural persons varies across datasets. This implies that for datasets where this identification is relatively more successful, the filter introduces a wedge between the total number of observations originally entailed in the dataset and the number of observations actually used for the record linkage purposes (after our filter is applied). One example are the datasets URS and BvD: in comparison, the URS dataset makes it easier to filter out natural persons, since they can be identified using the detailed economic sector information included in the URS. The BvD-data, on the other hand, does not include such a "natural-persons-indicator" consistently across time.[16)] As a result, we can expect to be relatively thourough in filtering out natural persons in the case of URS. Over all datasets, this natural-persons-filter likely has the highest impact for URS. We also see a more pronounced effect of the filter for BvD, to some extent for

13 This also alleviates a phenomenon often causing multiple assigments, which occurs when companies have typical german family names, such as "Hans Schmidt GmbH." Also, the effect on matching success can be expected to be small with regard to Bundesbank datasets, since Bundesbank analytical datasets for companies typically contain by design, if any, a very low number of observations on natural persons.
14 At least firm name and location information must be filled.
15 This applies mostly to AWMuS, which is the master database for foreign statistics. Note also that for the descriptive analyses in this report, MiDi entries were already limited to entities resident in Germany to begin with, i.e. to German mothers and German daughters of foreign mothers, which is why in Figure 1 we do not see a large effect of the filter applied for the MiDi dataset. The filter on resident companies also affects our BvD data, which includes a significant number of foreign entities for recent years, as well as RIAD and BAKIS data, which include foreign creditors.
16 In the BvD-Data available to us, such and indicator is only available for recent years

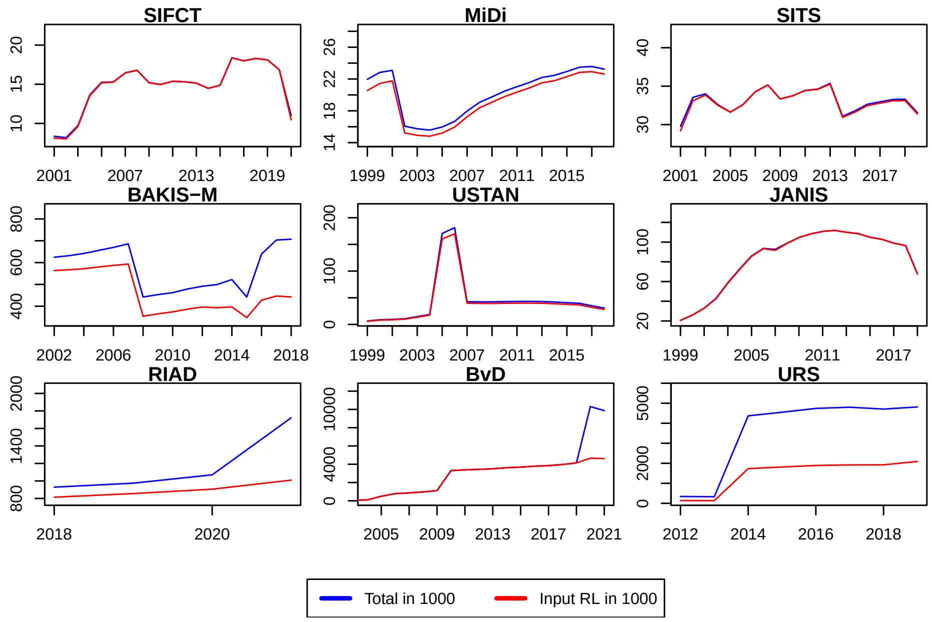

Figure 2: Unique entities in the input data over time

AWMuS and only a small effect for the other datsets.

2.6 Entities over Time

The number of companies for which information is collected based on the different legal mandates of the Bundesbank varies over time both within and across datasets. Further, due to the legal mandate and changes thereof over time, due to changes in firm characteristics, there is a certain level of attrition throughout the observation period. Data collection on some firms will be discontinued over time and some other firms will enter the data. This implies that the number of companies present in any given dataset is not constant over time. This is illustrated by Figure 2, "Unique entities in the input data over time". The blue lines in Figure 2 show for each dataset and each observation year the count of unique IDs, the red lines on the other hand visualize the effect of applying our filter described above on the number of unique IDs, separately for each year.[17)] Further, Figure 2 reveals that in some cases there is a more pronounced upward or downward shift in the number of IDs over time. In this section we take a closer look at these changes and provide some further clarification where applicable.

SIFCT

Between 2003-2004 and 2015-2016, respectively, we observe a significant increase in the number of companies recorded in the SIFCT dataset. In 2004 there are approximately 4000 more firms in the SIFCT dataset compared to 2003 and this is attributable to the successful integration of several firms that have missed their reporting obligations in earlier years.[18)] A similar increase in the number of firms between 2015-2016 is partly attributable to recovering several reporters,

17 Since only two datasets start before 1999 (USTAN and JANIS), the displayed time-spans are limited to begin in 1999.
18 2004 marks the establishment of the Foreign Trade Statistics Service Center of the Bundesbank, which informed all active and nonactive reporters to the SIFCT about this organisational change. This has helped numerous firms, considered as nonactive reporters, to comply again with their reporting obligations towards the Bundesbank.

who have missed their reporting requirements due to a change in the Foreign Trade and Payments Ordinance (Außenwirtschaftsverordnung) in 2013 and partly attributable to technical adjustments to reports submitted with the same reporting identifier. We also observe a drop in the number of unique IDs towards the end of the sample, which may be due to time lags in reporting and processing reports.

MiDi

There is a significant drop in the number of companies in MiDi in 2002, which is due to an increase in reporting thresholds in 2002. After 2002, we observe a moderate and sustained increase of German resident entities in the MiDi, which according to the documentation is likely due to the combination of nominally unchanged reporting thresholds plus inflation, as well as real growth of foreign direct investment during this time period.

SITS

There are no noteworthy historical changes, the fluctuation in the number of unique IDs over time corresponds to a regular turnover in the number of firms that engage in international trade in services. We also observe a drop in the number of unique IDs towards the end of the sample, which may be due to time lags in reporting and processing reports.

BAKIS-M

For the BAKIS-M data, we had to rely on different data deliveries, which, based on their content, can be categorized into 3 different categories: the first type of datasets were originally generated for an internal research project and cover the years 2002-2007. These datasets are, among other differences, characterized by a considerably larger absolute number of single proprietorships and limited information suitable to filter our private individuals. Also, the BAKIS-M data for 2002-2007 contain a considerably smaller absolute number of corporate legal forms. The drop in 2008 is therefore very likely to be mostly technical.[19)] The second type of dataset covers all other years, except 2015. The dataset for 2015 seems to be different, at least with regard to the share of missing information. We therefore presume the drop in 2015 to be very likely also be of a technical nature. Apart from such technical discontinuities, we presume the general upward trend of the case numbers to be attributable to the combination of nominally unchanged reporting thresholds, inflation, and also real growth of large loans.[20)]

USTAN

The spike in the number of companies around 2006 is likely attributable to the data migration from the predecessor of JALYS (i.e. the original master database "USTAN") into JALYS. In the course of this data migration, the time reference of every company record in USTAN, that was migrated to JALYS, seems to have been set to the year 2006 (the year in which the migration occurred), regardless of the original time reference information.

---

19 Although, since the drop occurs in 2008, part of it may also be attributable to the financial crisis that occured in that year. Since for this record linkage we only had access to the master data, but not to the amount data, we cannot verify this presumption.

20 Since we only had access to the master data, not the amount data of BAKIS-M, our options to verify these presumptions are limited.

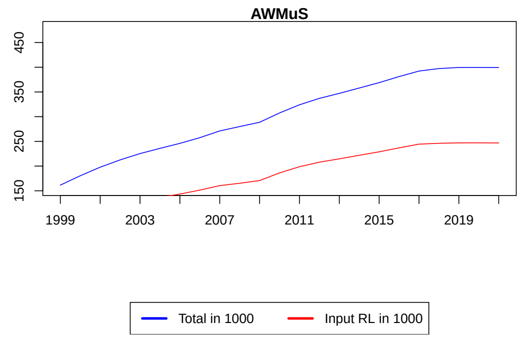

Figure 3: Unique entities in AWMuS over time

JANIS

We observe a drop in the number of unique records towards the end of the JANIS sample, which is due to time lags in reporting and processing the company reports.[21)].

RIAD

Mandatory reporting requirements for banks to the RIAD system were introduced in early 2018. Almost all time references are therefore larger than 2017.

AWMuS

The increasing number of records in AWMUS (Figure 3, "Unique entities in AWMuS over time") partly reflect the gradual increase over time in the number of companies that engage in foreign trade and therefore are required to report their activities to the Bundesbank. It is also known that records referring to inactive units are not always systematically detected and removed, therefore the steady increase is likely to be also partly due to inactive units remaining in the database.

BvD

Generally we observe a gradual increase in the number of uniqe IDs over time in the BvD dataset. This can be explained by, for example, improvements in record management, technical improvements in the data collection, and extanding the list of legal forms for which data was collected. Further, we observe a more remarkable shift in 2008 in the data. This corresponds to the time when the external data provider, Bureau van Dijk, extended its data aquisition, adding companies to the dataset that previously have not been given any credit rating from Creditreform, a German credit rating agency. Lastly, the decrease in the number of companies after 2012 is attributable to a change in the data collection related to an EU-Directive and its implementation in the German legal system (Kleinstkapitalgesellschaften / Bilanzierungsänderungsgesätz (MicroBilG), EU-Directive 2012/6/EU). This regulation allowes firms to store their balance sheet information in the database of the German Federal Gazette instead of publishing it in the Gazette. For firms specialized in data collection this had the important consequence that data stored in the database were subject to a fee whereas data published in the Gazette are publicly accessible. The increase in 2020 is due

21 As well as due to the practice of more recent reports being included before all reports for the recent years have been received.

to the fact that the BvD data available to us includes a significant number of foreign companies, natural persons, as well as company establishment entries for these later years. Of these, note that foreign companies and natural persons are filtered out for the purpose of our record linkage (see above), however for the establishment entities we opted to leave them in the data, since at least one Bundesbank dataset, notably the MiDi, by design also includes company establishments.[22)]

URS

Each year the Federal Statistical Office in Germany provides a snapshot of its business register to the Deutsche Bundesbank. Generally, the data transmitted to the Bundesbank contains information on companies that in a given reporting year not only contribute to the German GDP but their turnover or employment is above a minimum threshold that defines the company as "statistically relevant." The data transmitted for the 2014 reporting year also includes companies that entered the business register for statistical purposes before 2014. The data transmitted for the 2014 as well as the data transmitted for recent years includes also some companies that are not considered statistically relevant. This change in the data transmission from 2014 to 2015 and onwards and the effort to tag companies for which master data related information was recoded earlier than 2014, is responsible for the large difference in the number of unique company IDs between 2014 and the years before and after.

3 Linked Company Data

3.1 Linkage Process

The record linkage process is described in Doll, Gabor-Toth, & Schild (2021). After master data standardization, a reduced match candidate list is determined using a blocking algorithm. For this list of match candidates, comparison features are calculated.[23)] Available match information based on common IDs is consolidated to derive ground truth on matches / non-matches, a subset of which is then fed as training data into a supervised machine learning matching model. The model yields a match probability for each candidate pair, which is then consolidated with match information based on common IDs and exact comparisons of standardized names and addresses. The aim of the postprocessing is to reduce the share of probabilistic assignments as far as possible. On average, probabilistic matches account for about 5.1 %[24)] of all found matches.[25)] Overall, our record linkage process has consistently proven to yield stable and high quality results.

In addition to the matches provided by the record linkage of the RDSC (Doll, Gabor-Toth, & Schild, 2021), the department responsible for RIAD (S15) also provided matching tables for some ID relations. These matching tables were included in our final mapping tables "IDLINK" (see below). This is the case for matches between RIAD and USTAN, RIAD and JANIS, as well as between RIAD and BAKIS-M. For these relations, matching results by the RDSC where consolidated with the matching results by the department S15. For these relations, both S15 and the RDSC found matches that the respective other department did not find, due to different matching and data cleaning methods,

22 The jump in 2020 is also likely to be at least partly due to a set of preliminary BvD-IDs (identifiable by an Asterisk in the ID) being included in the more recent BvD data available, which constitute records the BvD has not yet matched and which therefore may be more likely to constitute duplicate records.
23 Such as different string distances on the firm names or geographical distance.
24 When considering all dataset relations between the datasets reported in this documentation.
25 Not surprisingly, there is a negative relationship between the share of probabilistic assignments and the share of common IDs found between datasets. Also, the share of probabilistic assignments decreases with the quality of the rest of the master data (especially names and addresses).

but also, to a much larger extent, due to different master data available to both departments. Both matching efforts were therefore consolidated in order to further improve the matching rate.[26]

3.2 Multiple ID Assignments

The ID-Tables include in some cases multiple assignments of ID-values: for example, an AWMUS identifier value may refer to two different JANIS ID values. Multiple ID-assignments may be caused by:

- different entity concepts: data sources may define entities differently. For example, next to legal entities, some data sources may assign IDs to establishments, which may lead to 1:m assignments.[27]
- implicitly different entity definitions: these are caused by how various master datasets address company restructurings (e.g. merger, split, aquisition). For example, two datasets may have different rules for dealing with mergers. One dataset may assign an entirely new ID for the newly created company, while another dataset may regard the created company as a successor of one of the companies that have merged, which it may express by giving the new company the ID value of one of the merged companies.
- ID-changes due to technical reasons: sometimes data providers assign new IDs to entities, for technical reasons (such as removal of duplicates). This problem is especially relevant for data sources with both a long time span and a set of different data deliveries, such as BvD and BAKIS-M, but also generally for data sources that contain a large share of duplicate records.

Providing data users with ID-tables that include multiple assignments leaves the researchers or analyst with the task of making sense of these multiple ID-assignments. On the other hand, removing multiple ID-assignments reduces the matching rate between two datasets, especially in the presence of long time spans and different entity concepts.

In the ID-mappingtables described in this data report, only multiple assignments that are based on exact identity of cleaned company names and place (city name) have been left in the tables. [28] This means that there are no multiple assignments that are based on either comparisons of common ID values (for example the trade register information or a borrower ID present in both databases) or that are based on probabilistic assignments.[29] Also, for the linkage tables described in this report, BvD-establishment-IDs as well as BvD-IDs marked as "preliminary"[30] that were part of a multiple assignment have been deleted.[31]

3.3 Linkage Results: IDLINK

The ID-linkage tables are two-column tables: one column for each ID. For each pair of company IDs, the ID linkage table provide answers to the question which ID-values represent the same real-world entity. These ID-linkage tables IDLINK are offered by the RDSC to internal analysts and, in an anonymized form, to researchers. The data product IDLINK is described by Gabor-Toth & Schild (2021).

26 In cases of contradictions between both matching results, preference was given to matches by S15.
27 Company establishments are included in AWMuS as well as in recent years of the BvD data.
28 For technical reasons, the ID-tables that were consolidated with the ID-tables by S15, i.e. "RIAD-USTAN", "RIAD-JANIS" and "RIAD-BAKISM", do not include any multiple assignments.
29 While multiple assignments based on identical ID-values for common identifiers have been discarded for the linkage tables described in this report (version 2021-2-6), they are available upon request as linkage table version 2021-2-5.
30 These are identifiable by an asterisk after the country code at the beginning of the ID.
31 Both are included in the linkage table version 2021-2-5, which is available upon request.

3.4 Overlap Analysis using IDLINK

Using the ID-linkage tables IDLINK (Gabor-Toth & Schild, 2021), we can generate overlaps between datasets, i.e. find units that are matchable and units that are only found in either one of the two datasets.

To make sense of the found overlaps, one must also consider the time dimension and the universes of the input data (hence the discussion of the time dimension and universes in the previous section). To see why this is the case, consider for example the master data of an imaginary company from USTAN for 2005 that was liquidated in 2009. Since the earliest master data in URS available for the Bundesbank stem from 2012, master data for this company will likely not exist in URS. Due to the existence of such companies with a theoretical probability to be matched being (close to) zero, our matching results expressed in percentages are "downward biased," at least when they are interpreted as a measure of (technical) matching success.

It is nevertheless informative to also take a look at the entire, or "pooled" overlaps between datasets, i.e. including all ID entries of both datasets, irrespective of their time reference or universe. Following the discussion of pooled overlaps, we will take a look at recent time periods, to see how many recent records from each dataset can be found in the respective other datasets ("overlaps for recent years"). To gain some understanding of the ID-linkage tables potential to generate linked data with a long time dimension, we would also like to know how matching success evolves when we vary the time reference of the data ("overlaps over time").[32] Finally, we look at "multilateral" overlaps, i.e. overlaps between more than two (here: three) datasets.

Pooled Overlaps

Table 3, "Matched entities, in 1000s, with entire time span for each dataset", on page 16, shows how many matches an analyst can expect when matching the above datasets where we take into account the full variation in the master data over the entire time span for every dataset.

The column and row "Total" in Table 3 shows the number of unique IDs for the respective dataset over the entire time span of this dataset. The rest of the columns contain the matching results. Values presented in each cell correspond to matching two datatsets at the intersection of a row and a column, with the first number being absolute match counts (in 1000), the second is in percentage terms. For example, around $36,000$ unique firm IDs from MiDi have been found in URS, that is approximately 69.4% of the MiDi firm IDs could be matched to an URS firm ID.

32 Note that, in contrast, we do not attempt to explicitly "harmonize" data universes, for example by looking at overlaps only for companies that we should expect to be very likely to be included in most datasets, such as large multinational companies. This will be a subject for a later investigation.

Table 3: Matched entities, in 1000s, with entire time span for each dataset

		AWMUS	BAKIS-M	BVD	JANIS	LEI	MIDI	RIAD	SIFCT	SITS	URS	USTAN
	Total	292	1189	7196	358	116	52	1012	73	131	2495	183
AWMUS	292		(105, 36.1)	(225, 77.0)	(79, 27.1)	(36, 12.3)	(52, 17.8)	(139, 47.5)	(73, 25.0)	(131, 44.7)	(172, 58.9)	(36, 12.2)
BAKIS-M	1189	(105, 8.8)		(477, 40.1)	(131, 11.0)	(55, 4.7)	(29, 2.5)	(348, 29.3)	(39, 3.3)	(58, 4.9)	(359, 30.2)	(57, 4.8)
BVD	7196	(267, 3.7)	(564, 7.8)		(397, 5.5)	(131, 1.8)	(59, 0.8)	(1114, 15.5)	(83, 1.2)	(136, 1.9)	(1974, 27.4)	(104, 1.4)
JANIS	358	(83, 23.1)	(137, 38.2)	(342, 95.5)		(32, 9.0)	(27, 7.6)	(184, 51.5)	(29, 8.1)	(48, 13.5)	(272, 75.9)	(71, 19.8)
LEI	116	(35, 30.5)	(55, 47.4)	(97, 83.6)	(30, 26.1)		(10, 8.9)	(108, 92.9)	(17, 14.9)	(23, 20.0)	(80, 68.7)	(14, 12.0)
MIDI	52	(52, 100.0)	(29, 56.1)	(47, 90.2)	(25, 48.6)	(10, 20.1)		(32, 61.1)	(25, 49.0)	(28, 54.2)	(36, 69.4)	(12, 22.4)
RIAD	1012	(139, 13.8)	(348, 34.4)	(953, 94.2)	(184, 18.2)	(109, 10.7)	(32, 3.2)		(45, 4.4)	(74, 7.3)	(797, 78.8)	(41, 4.1)
SIFCT	73	(73, 100.0)	(38, 52.1)	(63, 85.8)	(27, 36.3)	(17, 23.8)	(25, 34.8)	(44, 59.9)		(39, 53.7)	(48, 65.8)	(12, 16.7)
SITS	131	(131, 100.0)	(58, 44.3)	(108, 82.7)	(46, 34.9)	(23, 17.9)	(28, 21.5)	(73, 55.9)	(39, 30.1)		(90, 69.2)	(20, 15.3)
URS	2495	(171, 6.9)	(359, 14.4)	(1757, 70.4)	(263, 10.5)	(80, 3.2)	(36, 1.4)	(794, 31.8)	(48, 1.9)	(91, 3.6)		(66, 2.6)
USTAN	183	(36, 19.9)	(58, 31.5)	(90, 49.0)	(70, 38.1)	(14, 7.9)	(12, 6.7)	(41, 22.4)	(13, 7.1)	(21, 11.4)	(67, 36.5)	

[a] Absolute numbers are in 1000s. The second value in brackets is the percentage of the dataset's entire time span that has been successfully matched to the respective column dataset's entire timespan.

Note that ID-counts in Table 3 always refer to the unique number of IDs found for the ID in the respective row. Also note that ID counts may differ across the diagonal, for example, we find more BAKIS-M-entities in USTAN than USTAN-entities in BAKIS-M. This is due to the fact that IDLINK contains some multiple assigments of ID-values (m:m).[33)]

In the following we would like to highlight some relevant results. First, Table 3 shows that a considerable share of the frequently used research datasets can be found in at least one of the available large reference datasets, URS or BvD. The latter two datasets represent the most comprehensive collection of company master data in Germany. Another observation is that MiDi, SITS and SIFCT match 100% to AWMuS, which is not surprising since AMWuS serves as their common 'single-point-of-truth' master database. Third, the datasets tend to match a bit better to the master data of BvD than to the reference data of URS. This may be due to filtering rules applied by the Federal Statistical Office in Germany (Destatis) for determining which entries are relevant for statistical evaluation ("Auswertungsrelevanz").[34)] We also note that although both RIAD and BAKIS-M contain borrower level information, with RIAD having a lower reporting threshold than BAKIS-M, we were not able to match all companies in BAKIS-M to RIAD. This may be due to the fact that the BAKIS-M data has a different time span than RIAD (BAKIS starts earlier and our RIAD data covers later years). Alternatively, it may be attributable to different instruments being considered as loans in these two datasets.

Overlaps for Recent Years

In contrast to this pooled analysis of matching success in Table 3, which covers the entire time span of the datasets, Table 4 on page 18 shows the matching results *for the latest year* of each dataset (in rows) when matched to any other datasets' entire time span (pooled over all years, as shown in the columns). The total column in Table 4 now shows the unique number of entries for the most recent year and for each dataset (in 1000). We see that the matching results improve considerably. This is attributable to the fact that we disregard old master data from those datasets that cover a very long time period. These older and often outdated entries increase the size of a master dataset but have relatively little potential to be matched to master data from other datasets for which data only cover a few and especially recent years. For example, master data in MiDi from 1999 naturally have a lower matching rate to URS master data, for which the earliest entries in our sample stem from 2012.

33 The same is true for Table 4, "Matched entities, in 1000s, of each dataset's last year (rows) to each dataset's entire time span (columns)", on page 18, which is discussed below.

34 For example, according to the concept of "Auswertungsrelevanz", a company is filtered out from the business register if it does not have at least one employee for at least one month in a reporting year. The URS data available to us contains some records that are marked as not statistically relevant, and we kept them in, but the data available to us does not contain all not statistically relevant records that were originally in the data.

Table 4: Matched entities, in 1000s, of each dataset's last year (rows) to each dataset's entire time span (columns)

		AWMUS	BAKIS-M	BVD	JANIS	LEI	MIDI	RIAD	SIFCT	SITS	URS	USTAN
	Total	292	1189	7196	358	116	52	1012	73	131	2495	183
AWMUS'2021	247		(95, 38.4)	(196, 79.3)	(70, 28.4)	(34, 13.8)	(36, 14.8)	(133, 53.9)	(64, 25.9)	(114, 46.1)	(163, 65.9)	(31, 12.4)
BAKIS-M'2018	442	(66, 14.9)		(277, 62.7)	(89, 20.2)	(44, 10.0)	(19, 4.2)	(252, 56.9)	(26, 5.9)	(38, 8.7)	(242, 54.8)	(33, 7.6)
BVD'2021	4615	(249, 5.4)	(496, 10.7)		(369, 8.0)	(118, 2.5)	(56, 1.2)	(1062, 23.0)	(74, 1.6)	(128, 2.8)	(1809, 39.2)	(97, 2.1)
JANIS'2019	68	(24, 34.8)	(39, 58.1)	(67, 98.7)		(13, 19.4)	(9, 12.6)	(48, 71.0)	(10, 14.7)	(15, 22.9)	(63, 93.1)	(20, 29.2)
LEI'2018	116	(35, 30.5)	(55, 47.4)	(97, 83.6)	(30, 26.1)		(10, 8.9)	(108, 92.9)	(17, 14.9)	(23, 20.0)	(80, 68.7)	(14, 12.0)
MIDI'2018	23	(23, 100.0)	(16, 72.5)	(22, 99.0)	(14, 60.0)	(7, 30.8)		(20, 89.3)	(14, 63.2)	(14, 63.4)	(21, 94.4)	(6, 26.2)
RIAD'2021	1009	(138, 13.7)	(348, 34.5)	(951, 94.2)	(184, 18.3)	(108, 10.7)	(32, 3.2)		(44, 4.4)	(73, 7.3)	(795, 78.8)	(41, 4.1)
SIFCT'2021	10	(10, 100.0)	(7, 67.1)	(10, 92.9)	(4, 38.4)	(5, 48.0)	(5, 43.3)	(9, 84.0)		(7, 70.5)	(9, 81.6)	(2, 19.7)
SITS'2020	31	(31, 100.0)	(19, 59.9)	(29, 93.2)	(14, 45.6)	(11, 33.7)	(9, 29.7)	(26, 81.5)	(15, 48.1)		(28, 89.9)	(6, 20.4)
URS'2019	2091	(155, 7.4)	(328, 15.7)	(1538, 73.5)	(242, 11.6)	(77, 3.7)	(32, 1.5)	(770, 36.8)	(44, 2.1)	(82, 3.9)		(61, 2.9)
USTAN'2018	28	(15, 51.8)	(24, 85.4)	(28, 97.9)	(26, 93.8)	(10, 34.2)	(5, 17.8)	(24, 85.9)	(7, 23.1)	(10, 33.9)	(27, 95.1)	

[a] Absolute numbers are in 1000s. The second value in brackets is the percentage of the dataset's last year that has been successfully matched to the respective column dataset.

Overlaps Over Time

We would also like to understand if the matching success is fairly constant or changing over time, for each bilateral relation. This is useful information for analysts who are interested in linked data that covers a certain time span. We therefore take a look at how matching shares evolve over time. We discuss these results for the most frequently used datasets for analytical or research purposes: MiDi, JANIS, SITS, SIFCT and BAKIS-M.

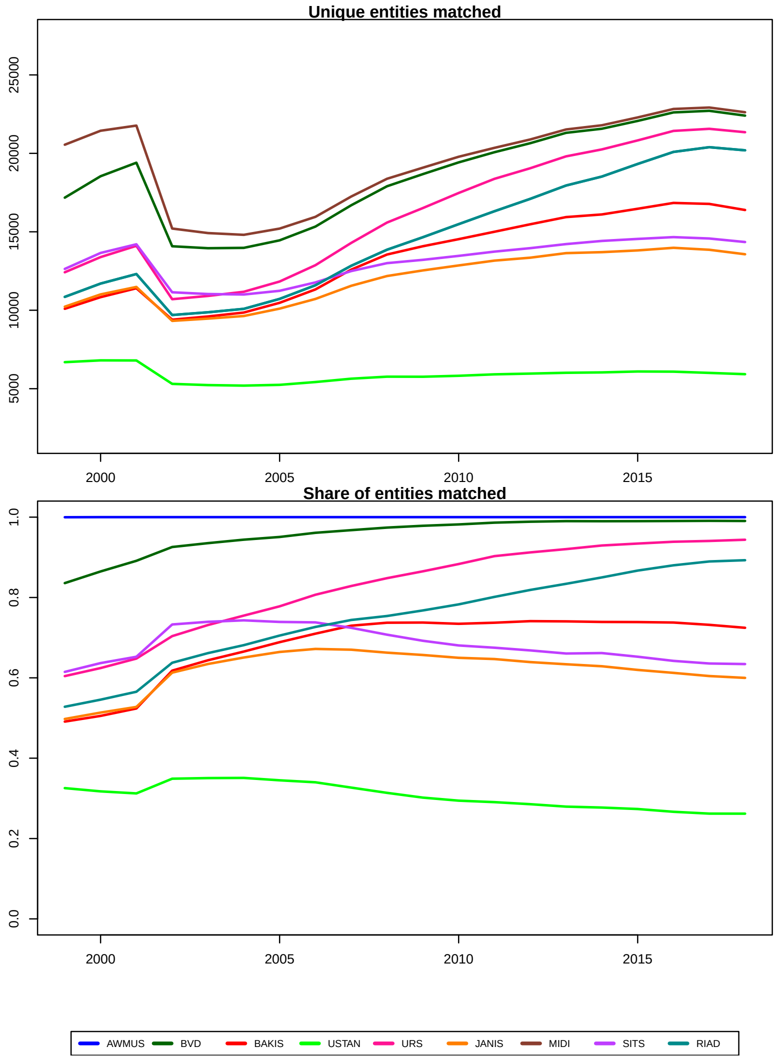

Figure 4: MiDi entities matched, by year

Figure 4, "MiDi entities matched, by year", zooms in on the MiDi dataset. The upper plot shows the total number of companies found in the respective datasets, by year. We see that the number of unique entities in the MiDi (German resident entities) drops in 2002. This is likely due to an increase in reporting thresholds as described in Blank, Lipponer, Schild, & Scholz (2020). Accordingly, the number of MiDi entries matched to other datasets also drops. Since the increase in reporting threshold made the average MiDi unit larger (and therefore both likelier to be included in other data and likelier to have well maintained, thus matchable master data), matching shares increase in 2002 for all datasets, as can be seen in the lower plot of Figure 4. For the following years, we observe that the number of MiDi units recovers steadily. This is likely due to a combination of largely nominally fixed reporting thresholds after 2002 in combination with inflation and growth in foreign direct investment activity over these years.

Most of the other datasets also grow in their size over time, which in combination causes increasing absolute match numbers. We also see that this increase is especially pronounced w.r.t. datasets that are only available for recent years. This is due to the fact that those contain only recent entries, which makes the common matchable set of datasets larger over time. When matching datasets with very different time spans or different master data availability, such as, for example, MiDi (1999-recent) and RIAD (only starting in 2019), this effect is most pronounced.

Such mismatch of time spans in linked data must be avoided, if possible, since it is likely to introduce survival bias. We therefore recommend using only time periods for linked data where the datasets involved both cover the respective timespan and an at least fairly stable number unique entities[35].

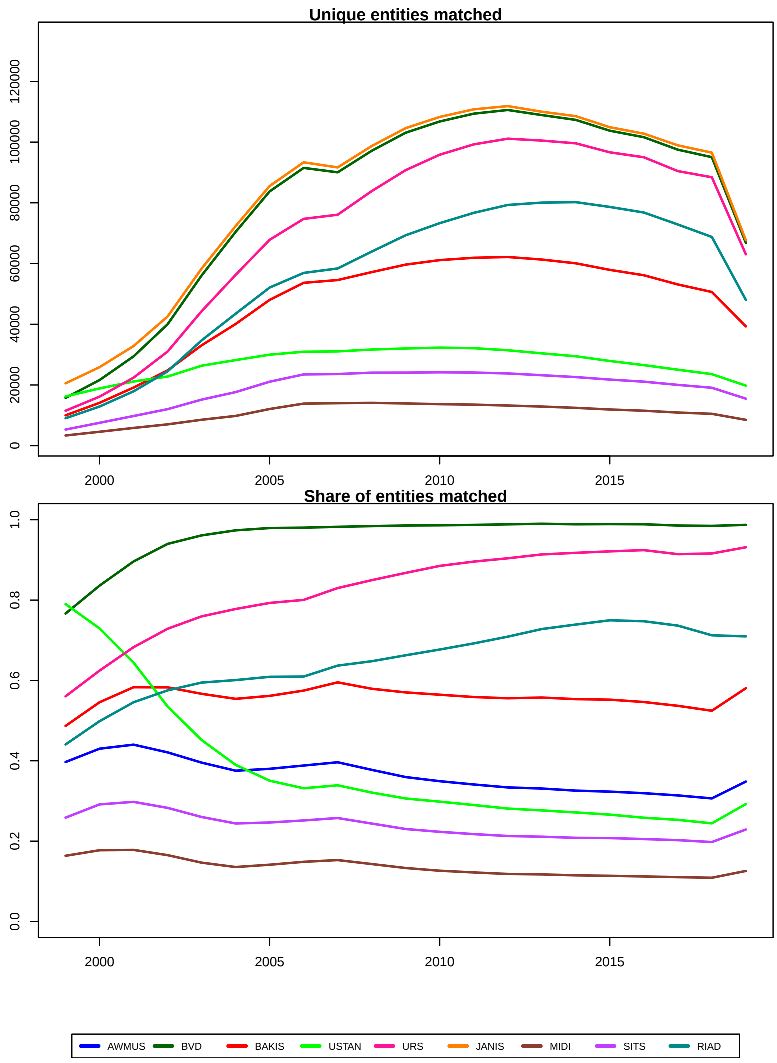

Figure 5: JANIS entities matched, by year

Figure 5 shows the matching results from the perspective of the JANIS dataset, by year. This figure

35 See Figure 2, "Unique entities in the input data over time", on page 10

shows the decreasing absolute and relative contribution of the USTAN dataset to the JANIS data over time, which is due to the decrease in balance sheet reporting following the introduction of the Euro (see Becker, Biewen, Schultz, & Weissbecker (2019a)), which is however in later years overcompensated by the availability of public balance sheet data.

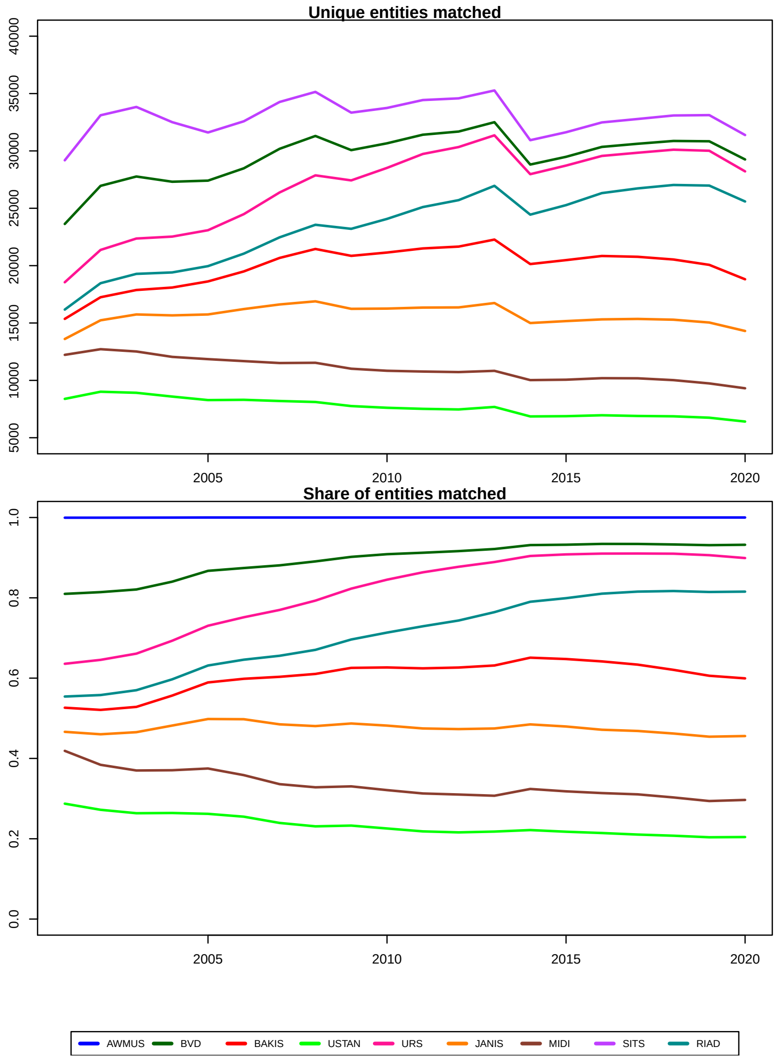

Figure 6: SITS entities matched, by year

Figure 6, "SITS entities matched, by year" shows the matching results from the perspective of the SITS dataset, by year. Here we also see the already discussed phenomenon of a steady increase in the share of companies matched to datasets that are either only available for recent years (RIAD, URS) or where there is a significant increase in the number of companies in the recent years (BvD). Since for these datasets a larger share of the observations is concentrated in more recent years, this makes the common matchable set for earlier years smaller. This is reflected in the rising share of SITS companies matched to URS, BvD and RIAD. Also, we see a slight drop in the total number of entities in 2014, which may be due to the introduction of a new standard for the calculation of the balance of payments statistics (change from BPM5 to BPM6). Finally, the share and also the absolute number of companies in SITS matched to MiDi, JANIS and USTAN decreases slightly over time. This is remarkable given that the size of MiDi and JANIS increases over time.[36] No clear explanation could be found for this pattern yet. It seems that it is unlikely due to record linkage success deteriorating for later years, since at least with the MiDi, the SITS has had the same master database over the time period in question, and therefore linkage should be close to complete, irrespective of the time reference.[37]

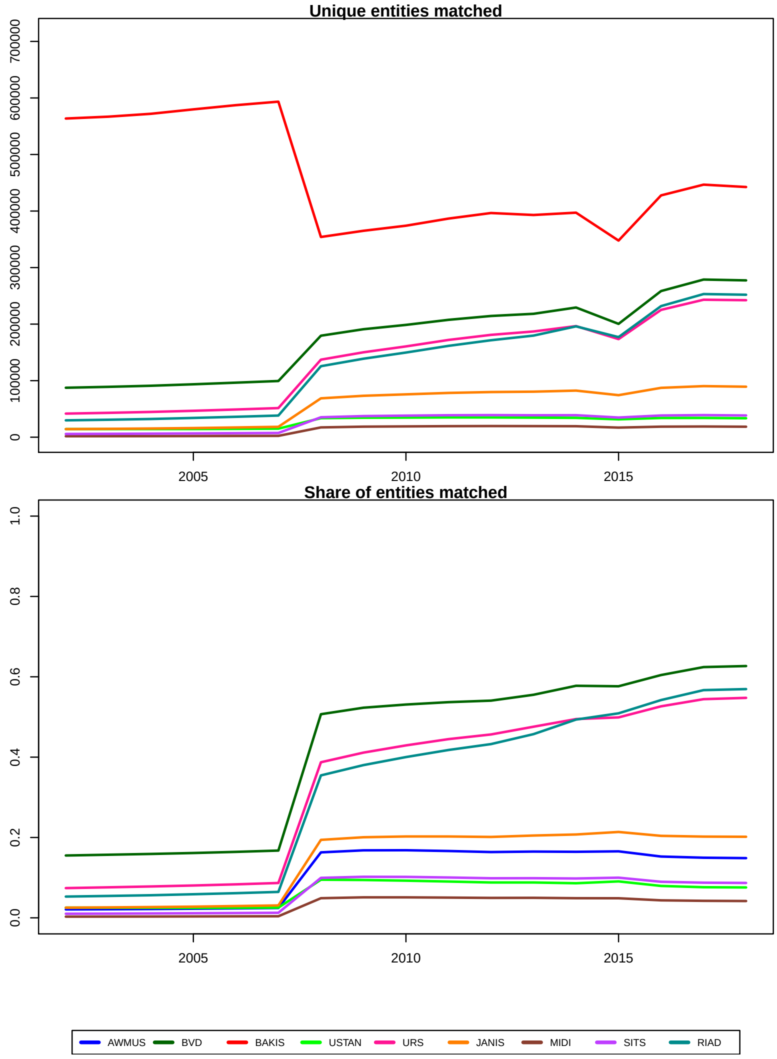

Figure 7: BAKIS-M entities matched, by year

Figure 7 shows the matching results from the perspective of the BAKIS-M dataset, by year. As discussed earlier, the size of this dataset decreases in 2008, which was made visible in Figure 2, "Unique entities in the input data over time", on page 10, but can also be seen here, by looking at the red line in the upper part of Figure 7. This is likely to a large part due to a much larger number of sole proprietorships included in the data for the years prior to 2008.[38] This interpretation fits

36 See Figure 2, "Unique entities in the input data over time", on page 10
37 We may speculate that the decreasing match share could be due to the distribution of the size of the companies in the SITS evolving differently than the size distributions in JANIS and MiDi over time. We have asked the department responsible for the SITS for an explanation. We will provide this explanation in a later version of this report, once available.
38 The number of entries with corporate legal forms increases in the later data, not compensating for the decrease in sole proprietorships.

to the fact that the number of BAKIS-M companies successfully matched to other datasets, which do not typically include a large share of sole proprietorships, increases in 2008. The relatively low matching rate to RIAD, URS and BvD remains unexplained for now.[39]

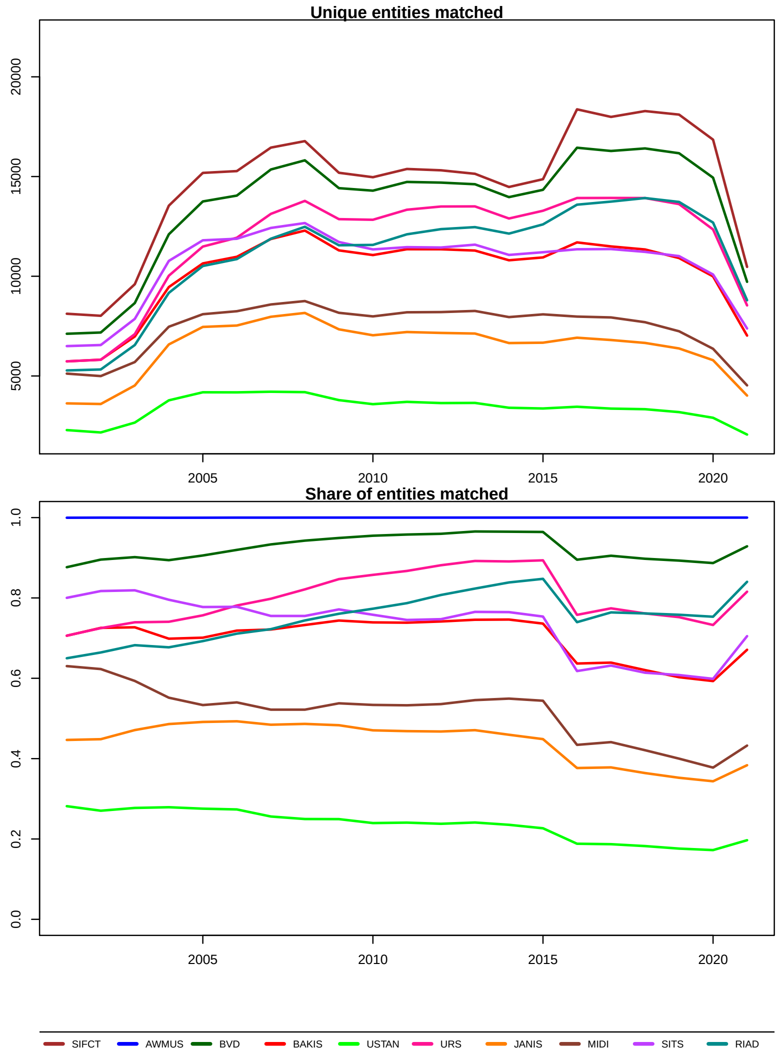

Figure 8: SIFCT entities matched, by year

Figure 8 shows the matching results from the perspective of the SIFCT dataset, by year. When the number of observations increases in SIFCT, so does the number of companies matched to other datasets. Looking at the share of companies matched, we observe a somewhat different pattern. The share of companies matched to URS, RIAD and BvD (where most of the observations concentrate in the recent few years) increases slightly over time until 2015, after which it suddenly decreases. Figure 8 shows a sudden increase in the number of companies captured in SIFCT after 2015, and a sudden decrease of matching shares w.r.t. the other datasets. We do not have a clear explanation for this pattern either. Again, it seems unlikely that this is due to the record linkage process, since the phenomenum also occurs w.r.t the MiDi, and since for the relation SIFTC to MiDi we do not need a record linkage, due to the two datasets sharing the same master database.

39 We suspect that it is due to insufficient filtering of company groups, natural persons or foreign companies in our BAKIS-M data.

We may again speculate: the drop in matching shares may be caused by increased coverage of SIFCT for entities that are less likely to be included in the other databases, for example smaller companies.[40]

Multilateral Pooled Overlaps

Finally, we take a look at multilateral overlaps. Such graphs not only visualize our earlier results but show multiple relations at the same time. To keep things simple, we limit the discussion to trilateral overlaps, and we only consider the frequently used datasets or the ones that are typically used together. This calculation, consistent with our previous statistics, is based on pooled data. This means that for each of the three datasets, we consider the dataset's entire time span.

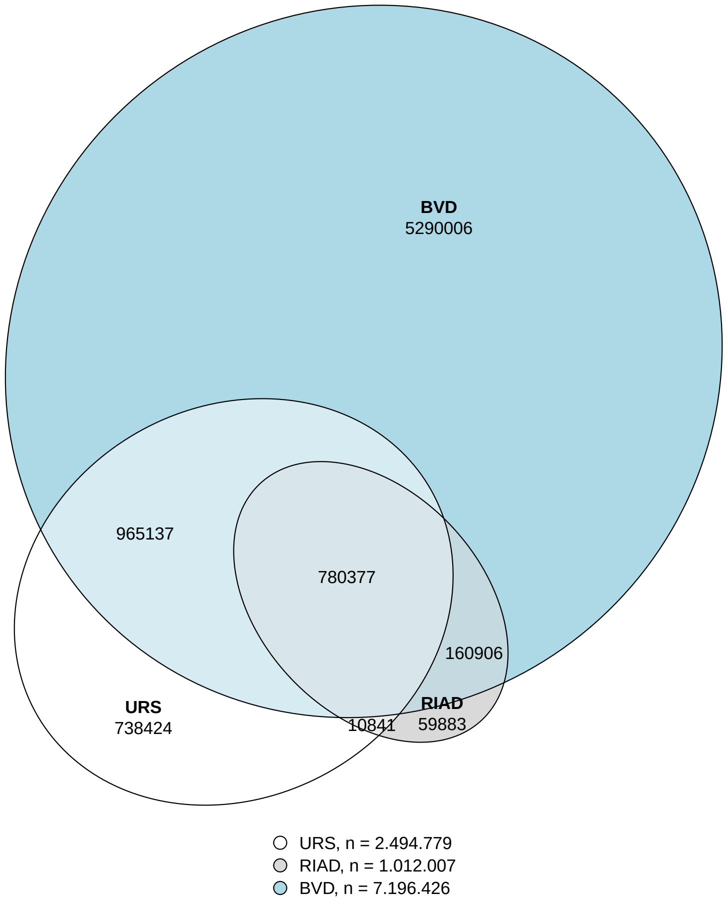

Figure 9: Matching overlap for BvD, URS and RIAD, (pooled)

In the first overlap graph, Figure 9, "Matching overlap for BvD, URS and RIAD, (pooled)" we see the matching overlaps for the three large reference datasets URS, BVD and RIAD. In the legend, the total number of records in each of the datasets is displayed. This graph further tells us, for example, that for 780.377 records in URS, we found a match to both other datasets (i.e. records found in all three datasets). It also tells us that for 965.137 records in URS, we found a match

40 We have asked the department responsible for the SIFCT for an explanation. We will provide this explanation in a later version of this report, one available.

in BVD, but not in RIAD. From these two numbers, we can calculate the entire bilateral overlap between URS and BVD by taking their sum, which yields a total of 1.745.514 records that we were able to match between URS and BVD, or, as a percentage, 70% of URS. Also, we see the number of records that do not match to any of the other datasets: in the case of URS, this amounts to 738.424 records, or, in percentage terms, 29,6% of all URS-records that were available for the record linkage.

Next to being the master database for the relatively new credit register AnaCredit, RIAD is also the central master database for several purposes within the Bundesbank and in the ESCB. Note that for RIAD, only 59.883 of 1.012.007 do not match to either URS or BVD, or, in percentage terms, 5,9%. Also, even with RIAD starting only in 2019, we were able to match most of the companies currently in RIAD (77,1%) to both URS and BvD. The large share of BvD companies not found in URS (75,7%) is partly attributable to the significant number of observations categorized as natural persons that we were able to filter out from URS compared to BvD, but, more importantly, to the much longer time span of the BvD data.

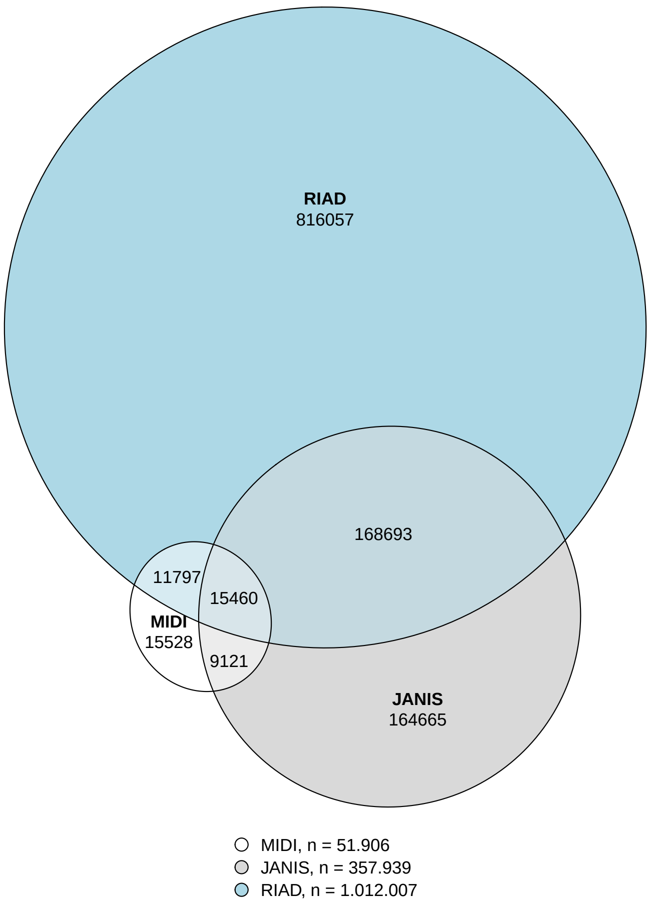

Figure 10: Matching overlap for MiDi, JANIS and RIAD, (pooled)

In Figure 10 we present the matching overlap for two prominent research datasets, MIDI and JANIS, with RIAD.[41] Apart from the size difference (MiDi and JANIS have a much smaller universe than RIAD), we note that in this pooled view, only a little more than half of the records in MIDI

41 This is of interest for analysts who would like to use AnaCredit jointly with either of these two research datasets.

(52,5%) and JANIS (51,4%) are matched to RIAD/AnaCredit. This is however mostly due to MiDi and JANIS, other than RIAD, comprising historic data: if we only look at recent entries, a much larger share of MiDi and JANIS is matched to RIAD.[42)]

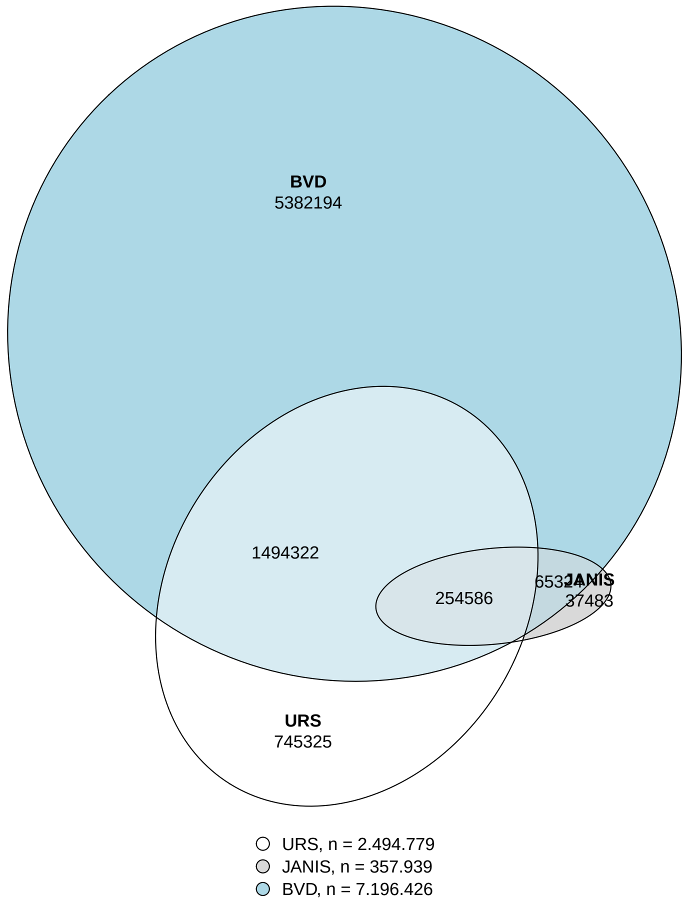

Figure 11: Matching overlap for BvD, URS and JANIS, (pooled)

Figure 11 shows the matching overlap for the three large reference datasets BvD, URS and JANIS. While URS is not available for research and analytics at Deutsche Bundesbank, it serves as a reference dataset that aims to comprise the entirety of statistically relevant companies located in Germany. BvD, on the other hand, also includes companies that are listed but would not necessarily be considered statistically relevant by the German Statistical Office. In addition, when comparing the size of the datasets as they are pictured in the graph, one has to consider that this is the pooled data avalailable for the record linkage: our BvD and JANIS data go back much further than the URS data (see section "Entities over Time").

42 To see this, compare the matching rates of MiDi and JANIS to RIAD in Table 4, "Matched entities, in 1000s, of each dataset's last year (rows) to each dataset's entire time span (columns)", on page 18.

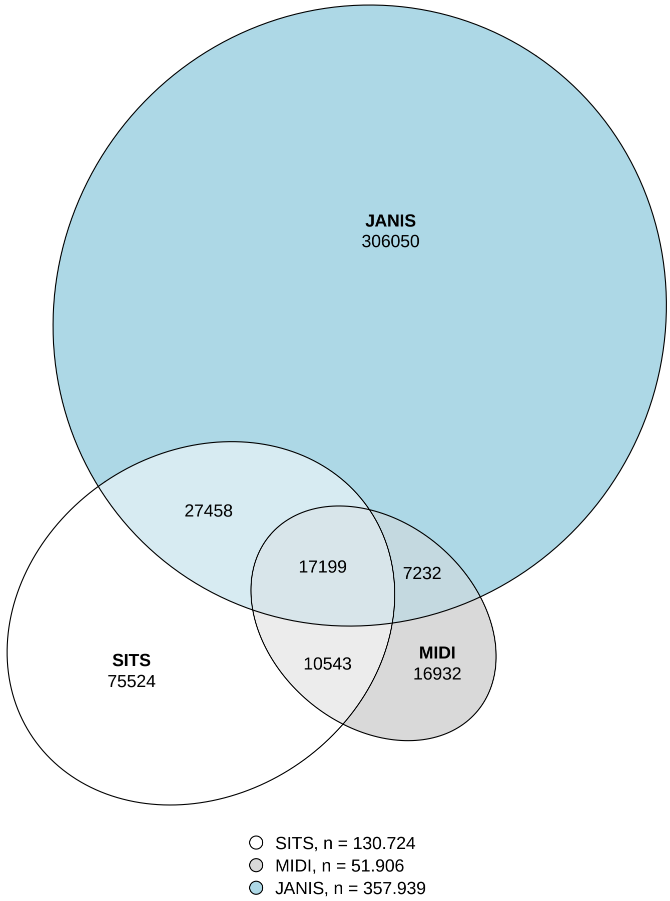

Figure 12: Matching overlap for JANIS, MiDi and SITS, (pooled)

Figure 12 shows the overlap between JANIS, MiDi and SITS. This is a combination of research datasets interesting for researchers who want to analyze services trade by multinational companies. Since neither SITS nor MiDi contain information on the German parent companies, especially regarding balance sheet information, the JANIS is often used together with these datasets. The graph shows that not every company involved in international services trade is part of a multinational company in the sense of the MiDi reporting thresholds. It also confirms that not all entities resident in Germany that are part of a multinational company show up in services trade statistics; this may be either because they do not engage much in services trade or because services trade is reported by a different german based entity within the multinational company. Lastly, a large share of entities in JANIS cannot be found in SITS or MiDi. This is plausible, given that many smaller and medium sized companies do not publish complete balance sheets that include profit and loss accounts, which is a requirement in order to be included in the JANIS database.

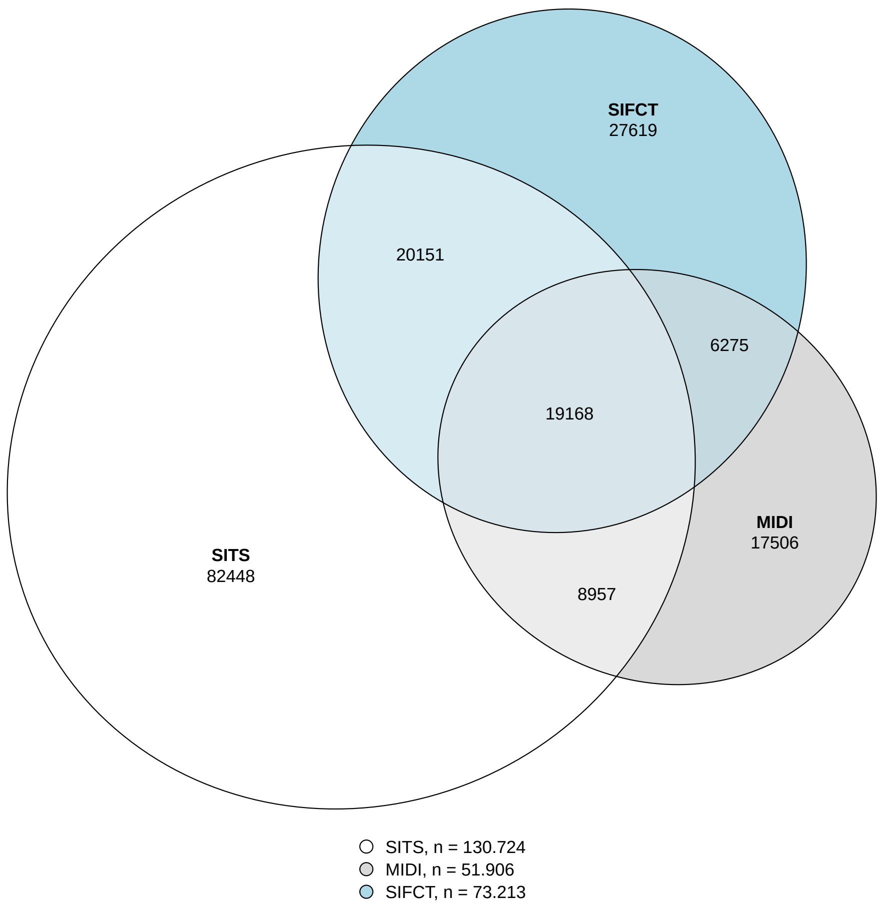

Figure 13: Matching overlap for MiDi, SIFCT, and SITS, (pooled)

Figure 13 focuses on the datasets relevant for international trade statistics, MiDi, SITS and SIFCT. This graph shows that a large number of companies are active in multiple forms of international trade. There is, however, also a significant share of companies that is only active in one form of international trade. Note again that these three datasets share the same master data (AWMUS), therefore the overlap graph in Figure 13 can be expected to come close to the real data universe overlap.

Finally, Figure 14, "Matching overlap for BAKIS-M, JANIS and RIAD, (pooled)", on page 29, presents the corresponding multilateral overlap for BAKIS-M, RIAD and JANIS. Although BAKIS-M and RIAD both contain borrower level information, their reporting threshold is different, with the one for RIAD (AnaCredit) being lower. This explains why companies from RIAD are not necessarily also in BAKIS-M. Further, since both BAKIS-M and JANIS cover a much longer time span compared to RIAD, a large share of the companies in BAKIS-M and JANIS will not be present in RIAD.

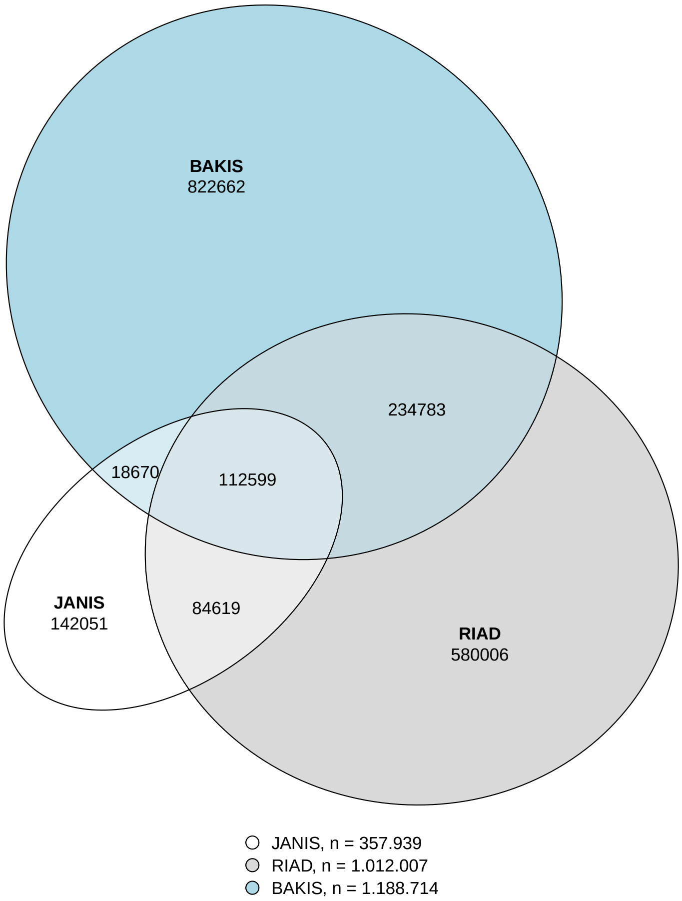

Figure 14: Matching overlap for BAKIS-M, JANIS and RIAD, (pooled)

4 Conclusion

The Deutsche Bundesbank collects company microdata for various purposes. The Research Data and Service Center offers company data derived from these company databases. These analytical and research datasets therefore originate from databases that were designed for different scopes, looking at only partly overlapping sets of companies. Consequently, the company research datasets contain complementing information for many companies, but not always on the same companies.

To maximize their analytical and research potential, the Research Data and Service Center links company research datasets. The linkage results show substantial overlaps, but not complete overlaps. To make sense of the overlaps realized by the record linkage of the RDSC, i.e. to gain an impression how closely they correspond to the true overlap between the research datasets, it is helpful to understand the quality of the matching process, which is therefore described by Doll, Gabor-Toth, & Schild (2021). Secondly, understanding the universe and time dimension of these datasets is necessary to better understand what is possible or not possible when linking these data. This is the purpose of the present data report.

The report therefores starts by describing various aspects of the datasets involved in this record linkage. We then present and disentangle our overall matching results. We show that a large share of the research datasets can be found in BvD and URS, the largest master datasets available to us. We are also able to successfully match RIAD with several Bundesbank research datasets. Further, our analysis makes clear how the datasets' time span greatly influences the success of the linkage: the share of companies matched between two datasets varies over time and often it reflects the underlying pattern in the datasets involved. The greater the common time overlap between two datasets, the higher is the share of successfully matched companies. To facilitate the understanding of these matching shares and the number of companies matched, we present our results from multiple angles and from the perspective of different datasets: pooled versus focusing only on the most recent year available, bilateral versus trilateral relationships, considering versus discarding the time dimension.

For future research, to further understand overlaps and how closely they correspond to true overlaps, we see three principal approaches: first, a systematic, manual research and analysis of a number of non-matched companies in the data could be worthwhile. Second, universes, and how they change over time, could be further explored based on information that is included in some of the datasets. For example, changing (diverging or converging) firm size distributions, or changing economic sector compositions, could explain some of the change in overlaps over time. Lastly, record linkage success could be investigated with a focus on the time dimension as well, for example by looking at changing prevalences of various entity identifying variables.

References

Alves-Werb, G., Krodel, T., Lassen, F., Orben, J., Schild, C.-J., & Schäfer, M. (2020). AnaCredit - German Part, (forthcoming). Data Report, Research Data and Service Centre, Deutsche Bundesbank.

Becker, T., Biewen, E., Schultz, S., & Weissbecker, M. (2019a). Individual Financial Statements of Non-financial Firms (JANIS) 1997-2017. Data Report 2019-10, Research Data and Service Centre, Deutsche Bundesbank.

Becker, T., Biewen, E., Schultz, S., & Weissbecker, M. (2019b). Corporate Balance Sheets (Ustan) 1987-2017. Data Report 2017-02, Research Data and Service Centre, Deutsche Bundesbank.

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

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

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

Biewen, E., & Lohner, S. (2019). Statistics on International Trade in Services (SITS) 2001-2018. Data Report 2019-07, Research Data and Service Centre, Deutsche Bundesbank.

Biewen, E., & Stahl, H. (2020). Statistics on International Financial and Capital Transactions (SIFCT). Data Report 2020-07, Research Data and Service Centre, Deutsche Bundesbank.

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

DESTATIS. (2019). Unternehmensregistersystem - Qualitätsbericht. Wiesbaden: Statistisches Bundesamt.

Doll, H., Gabor-Toth, E., & Schild, C.-J. (2021). Linking Deutsche Bundesbank Company Data. Technical Report 2021-05, Research Data and Service Centre, Deutsche Bundesbank.

ECB RIAD Team. (2019). RIAD Reference Manual, (RHN19.72-2). European Central Bank.

Gabor-Toth, E., & Schild, C.-J. (2021). Company (ID) Linktables - (IDLINK). Data Report 2021-22, Research Data and Service Centre, Deutsche Bundesbank.

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

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