DEUTSCHE BUNDESBANK
EUROSYSTEM

# Company ID Linktables (IDLINK)
## Data Report 2021-22

Data available from 1987-01-01 to 2021-03-01
Version: 2021-2-6
DOI: 10.12757/BBk.IDLINK.198701-202103

Deutsche Bundesbank, Research Data and Service Centre

Eniko Gábor-Tóth
Christopher-Johannes Schild

Research Data and Service Centre

## Abstract

This data report describes the research dataset “IDLINK” the company ID-linkage tables produced by the RDSC company data record linkage[1)], using a structured metadata schema.[2)]

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

**Version:** 2021-2-6

**DOI:** 10.12757/BBk.IDLINK.198701-202103

**Citation:** Eniko Gábor-Tóth, Christopher-Johannes Schild, Company ID Linktables - IDLINK, Data Report 2021-22– Version2021-2-6. Deutsche Bundesbank, Research Data and Service Centre.

1 The department responsible for RIAD (S15) provided additional ID-mappings between RIAD and USTAN, RIAD and JANIS, and between RIAD and BAKIS-M. These tables also entered the ID-linkage tables in IDLINK: they were consolidated with the matching result from the RDSC record linkage in order to further improve the matching rate.

2 We thank Katja Ziprik and Katharina Muno for providing additional match information produced by the RIAD Team of Deutsche Bundesbank. The metadata scheme is derived from the “Data Documentation Initiative” (DDI, http://www.ddialliance.org).

# Contents

# 1 Dataset Description

## 1.1 Overview and Identification

The "IDLINK" dataset comprises 43 ID linkage tables. Each of these 43 tables allows for linking two research (analytical) datasets. Through the pairwise linkage of BVD, MiDi, URS, USTAN, JANIS, BAKIS-M, LEI, RIAD (AnaCredit), SIFCT, SITS, "IDLINK" allows for jointly evaluating these datasets for various time periods.

The individual ID linkage tables are two-column tables containing a different company identifier in each column, resulting in a table of ID value pairs, showing which ID values of the two IDs refer to the same real-world company entity. That is, the ID linkage tables contain the list of native ID pairs (or anonymized versions thereof) that form a match between two datasets. The need for these ID linkage tables originates from the fact that company data is often held in separate databases, which use different company identifiers. The ID linkage tables allow linking different company datasets.

The tables are produced by the RDSC through the use of current record linkage techniques, which, among other methods, include comprehensive data cleaning, matching based on common external IDs, as well as name and place based matching, which includes probabilisic matching using supervised machine learning. The technical properties of this record linkage process are described by Doll, Gabor-Toth, & Schild (2021).

The department responsible for RIAD (S15) provided additional ID-mappings between RIAD and USTAN, RIAD and JANIS, and between RIAD and BAKIS-M. These tables also entered the ID-linkage tables in IDLINK: they were consolidated with the matching result from the RDSC record linkage in order to further improve the matching rate.

The ID value pairs included in these linkage table define overlaps between the different company datasets. These overlaps are described by Gabor-Toth & Schild (2021).

## 1.2 Dataset Scope and Coverage

### Legal Framework

The Research Data and Service Centre (RDSC) is mandated by the Deutsche Bundesbank to use internal company master data for linkage purposes, to provide linked company data to internal analysts and anonymized linked company data to internal and external researchers.

### Unit of Analysis

The unit of analysis is a single company. Generally this refers to legal entities. In each dataset to be linked, entities are represented by their dataset-specific native identifier.

### Time Periods

The time periods covered by the data depend on the time periods of the input data sources. The earliest time references for ID values in the input data stem from 1979, the latest stem from 2019. For details see Gabor-Toth & Schild (2021)

### Geographic Coverage

Germany

### Universe

The universes of the input datasets of our record linkage define the universe of our ID linkage tables. This universe covers all companies that appear at least once in any of the master datasets that have entered the record linkage. For details on the datasets' universes, we refer to the respective datasets' official documentation. For an overview on these datasets' universes and for their respective documentations see Gabor-Toth & Schild (2021)

### Historical Changes

The input datasets' universes change over time, for example due to changes of reporting thresholds, or quality of data historization. Data integration for historic company data at Deutsche Bundesbank is in its infancy, understanding the different company databases and their universes and how these universes change over time is one of the challenges. For an attempt to summarize the current state of knowledge on these matters see Gabor-Toth & Schild (2021)

## 1.3 Data Collection

## 1.4 Data Appraisal

### Quality Checks

Quality checks are implemented at several points during the record linkage process. For example, before entering the record linkage, certain quality checks are undertaken for the input data, such as examining the filling ratios for positions that are relevant for linking the datasets. The ID-linkage tables are analyzed for match prediction precision and recall / match coverage, see Doll, Gabor-Toth, & Schild (2021). Match coverage is further analyzed by Gabor-Toth & Schild (2021), with a focus on discussing the plausibility of the implied sizes of the data universe overlaps.

### Data Editing

Before entering the record linkage that produced the ID-linkage tables described in this report, the input data is standardized by the data specialist of the RDSC responsible for the dataset. Standardization occurs according to the data standard defined by AnaCredit RIAD, currently v2.1.[3)].

### Data Anonymization

Before the ID linkage tables are provided to researchers, all external company identifiers are anonymized using a secure hash algorithm (SHA256) as recommended by the German Federal Institute for Information Security (Bundesamt für Sicherheit in der Informationstechnik, 2020).[4)]

## 1.5 Data Accessibility

Data from external sources such as BvD can only be accessed if the researcher has a licence with the data provider that allows the use of their data. Some datasets that have no research dataset counterpart cannot be accessed by external users, such as the case for the URS dataset, which we are allowed to use only for statistical purposes such as data linkage, but not for generating research data.

### Research Proposal Conditions

A research proposal is checked for feasibility of the research project given the research data, i.e. the suitability of the data to answer the research questions raised by the proposal. The research project must be of public interest, that is without commercial goals.

### Institutional Access Conditions

The researcher must be affiliated with a research institution that clearly has a scientific, noncommercial agenda.

### Contact

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

3 https://www.bundesbank.de/de/service/meldewesen/bankenstatistik/formate-xml/formate-xml--611846
4 To enable the researcher to use the ID-mappingtables to link different datasets, the IDs contained in the research data are hashed using the same hash algorithm.

### Deposit Requirements

The researcher must sign a confidentiality agreement and a contract with the Deutsche Bundesbank. To use microdata available in BAKIS-M a separate contract needs to be signed between the research institute that the researcher is affiliated with and the Deutsche Bundesbank

### 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:

Eniko Gábor-Tóth, Christopher-Johannes Schild, Company ID Linktables - IDLINK, Data Report 2021-22– Version2021-2-6. Deutsche Bundesbank, Research Data and Service Centre.

## 1.6 Files Description

The files contain ID value pairs, showing which ID values of the two IDs refer to the same real-world company entity. The 28 files in Table 1, "Files - ID LINK 'not anomymized'", on page 8, contain non-anonymized identifiers and are only available to internal analysts. The 15 files in Table 2, "Files - ID LINK 'anomymized'", on page 8, only contain anonymized identifiers and are available to researchers.

### File Structure

The files contain 2 columns (two IDs).

### Type of Files

ASCII

### Data Formats

comma-delimited

Table 1: Files - ID LINK 'not anomymized'

| Filename | IDs | Datasets |
|---|---|---|
| *orig_v2021-2-6_BVD_CD_AWMUS_CD* | BVD_CD - AWMUS_CD | BVD - AWMuS |
| *orig_v2021-2-6_DE_BAKISN_CD_AWMUS_CD* | DE_BAKISN_CD - AWMUS_CD | BAKIS-M - AWMuS |
| *orig_v2021-2-6_DE_BAKISN_CD_BVD_CD* | DE_BAKISN_CD - BVD_CD | BAKIS-M - BVD |
| *orig_v2021-2-6_DE_DESTATIS_CD_STBL_AWMUS_CD* | DE_DESTATIS_CD_STBL - AWMUS_CD | URS - AWMuS |
| *orig_v2021-2-6_DE_DESTATIS_CD_STBL_BVD_CD* | DE_DESTATIS_CD_STBL - BVD_CD | URS - BVD |
| *orig_v2021-2-6_DE_DESTATIS_CD_STBL_DE_BAKISN_CD* | DE_DESTATIS_CD_STBL - DE_BAKISN_CD | URS - BAKIS-M |
| *orig_v2021-2-6_ENTTY_RIAD_CD_AWMUS_CD* | ENTTY_RIAD_CD - AWMUS_CD | AnaCredit - AWMuS |
| *orig_v2021-2-6_ENTTY_RIAD_CD_BVD_CD* | ENTTY_RIAD_CD - BVD_CD | AnaCredit - BVD |
| *orig_v2021-2-6_ENTTY_RIAD_CD_DE_BAKISN_CD* | ENTTY_RIAD_CD - DE_BAKISN_CD | AnaCredit - BAKIS-M |
| *orig_v2021-2-6_ENTTY_RIAD_CD_DE_DESTATIS_CD_STBL* | ENTTY_RIAD_CD - DE_DESTATIS_CD_STBL | AnaCredit - URS |
| *orig_v2021-2-6_JANIS_CD_AWMUS_CD* | JANIS_CD - AWMUS_CD | JANIS - AWMuS |
| *orig_v2021-2-6_JANIS_CD_BVD_CD* | JANIS_CD - BVD_CD | JANIS - BVD |
| *orig_v2021-2-6_JANIS_CD_DE_BAKISN_CD* | JANIS_CD - DE_BAKISN_CD | JANIS - BAKIS-M |
| *orig_v2021-2-6_JANIS_CD_DE_DESTATIS_CD_STBL* | JANIS_CD - DE_DESTATIS_CD_STBL | JANIS - URS |
| *orig_v2021-2-6_JANIS_CD_ENTTY_RIAD_CD* | JANIS_CD - ENTTY_RIAD_CD | JANIS - AnaCredit |
| *orig_v2021-2-6_LEI_AWMUS_CD* | LEI - AWMUS_CD | LEI - AWMuS |
| *orig_v2021-2-6_LEI_BVD_CD* | LEI - BVD_CD | LEI - BVD |
| *orig_v2021-2-6_LEI_DE_BAKISN_CD* | LEI - DE_BAKISN_CD | LEI - BAKIS-M |
| *orig_v2021-2-6_LEI_DE_DESTATIS_CD_STBL* | LEI - DE_DESTATIS_CD_STBL | LEI - URS |
| *orig_v2021-2-6_LEI_ENTTY_RIAD_CD* | LEI - ENTTY_RIAD_CD | LEI - AnaCredit |
| *orig_v2021-2-6_LEI_JANIS_CD* | LEI - JANIS_CD | LEI - JANIS |
| *orig_v2021-2-6_USTAN_CD_AWMUS_CD* | USTAN_CD - AWMUS_CD | USTAN - AWMuS |
| *orig_v2021-2-6_USTAN_CD_BVD_CD* | USTAN_CD - BVD_CD | USTAN - BVD |
| *orig_v2021-2-6_USTAN_CD_DE_BAKISN_CD* | USTAN_CD - DE_BAKISN_CD | USTAN - BAKIS-M |
| *orig_v2021-2-6_USTAN_CD_DE_DESTATIS_CD_STBL* | USTAN_CD - DE_DESTATIS_CD_STBL | USTAN - URS |
| *orig_v2021-2-6_USTAN_CD_ENTTY_RIAD_CD* | USTAN_CD - ENTTY_RIAD_CD | USTAN - AnaCredit |
| *orig_v2021-2-6_USTAN_CD_JANIS_CD* | USTAN_CD - JANIS_CD | USTAN - JANIS |
| *orig_v2021-2-6_USTAN_CD_LEI* | USTAN_CD - LEI | USTAN - LEI |

[a] AWMuS holds the master data for MiDi, SITS and SIFCT. Therefore, through AWMuS, datasets can be linked to the three Bundesbank research datasets related to international trade statistics.

Table 2: Files - ID LINK 'anomymized'

| Filename | IDs | Datasets |
|---|---|---|
| *anon_v2021-2-6_BBK_HM_CO_ID1_AWMUS_CD* | BBK_HM_CO_ID1 - AWMUS_CD | JANIS - AWMuS |
| *anon_v2021-2-6_BBK_HM_CO_ID1_BVD_CD* | BBK_HM_CO_ID1 - anonym_BVD_CD | JANIS - BVD |
| *anon_v2021-2-6_BBK_HM_CO_ID1_ENTTY_RIAD_CD* | BBK_HM_CO_ID1 - anonym_ENTTY_RIAD_CD | JANIS - AnaCredit |
| *anon_v2021-2-6_BVD_CD_AWMUS_CD* | anonym_BVD_CD - AWMUS_CD | BVD - AWMuS |
| *anon_v2021-2-6_ENTTY_RIAD_CD_AWMUS_CD* | anonym_ENTTY_RIAD_CD - AWMUS_CD | AnaCredit - AWMuS |
| *anon_v2021-2-6_ENTTY_RIAD_CD_BVD_CD* | anonym_ENTTY_RIAD_CD - anonym_BVD_CD | AnaCredit - BVD |
| *anon_v2021-2-6_LEI_AWMUS_CD* | anonym_LEI - AWMUS_CD | LEI - AWMuS |
| *anon_v2021-2-6_LEI_BBK_HM_CO_ID1* | anonym_LEI - BBK_HM_CO_ID1 | LEI - JANIS |
| *anon_v2021-2-6_LEI_BVD_CD* | anonym_LEI - anonym_BVD_CD | LEI - BVD |
| *anon_v2021-2-6_LEI_ENTTY_RIAD_CD* | anonym_LEI - anonym_ENTTY_RIAD_CD | LEI - AnaCredit |
| *anon_v2021-2-6_USTAN_CD_AWMUS_CD* | anonym_USTAN_CD - AWMUS_CD | USTAN - AWMuS |
| *anon_v2021-2-6_USTAN_CD_BBK_HM_CO_ID1* | anonym_USTAN_CD - BBK_HM_CO_ID1 | USTAN - JANIS |
| *anon_v2021-2-6_USTAN_CD_BVD_CD* | anonym_USTAN_CD - anonym_BVD_CD | USTAN - BVD |
| *anon_v2021-2-6_USTAN_CD_ENTTY_RIAD_CD* | anonym_USTAN_CD - anonym_ENTTY_RIAD_CD | USTAN - AnaCredit |
| *anon_v2021-2-6_USTAN_CD_LEI* | anonym_USTAN_CD - anonym_LEI | USTAN - LEI |

[a] AWMuS holds the master data for MiDi, SITS and SIFCT. Therefore, through AWMuS, datasets can be linked to the three Bundesbank research datasets related to international trade statistics.

# 2 Description of Variables

## 2.1 Overview of Variables

| Name | Label |
|---|---|
| AWMUS_CD | AWMuS identifier. Anonymous. Original name: MLD_NR |
| BBK_H_CO_ID1 | JANIS identifier (anonymized) |
| BVD_CD | Bureau van Dijk identifier. Original name: bvdid |
| DE_BAKISN_CD | Borrower ID (”Nehmernummer”) |
| DE_DESTATIS_CD_STBL | Destatis business register ID. Original name: WE_ID_ALT |
| ENTTY_RIAD_CD | RIAD identifier. |
| LEI | Legal entity identifier |
| USTAN_CD | USTAN identifier. Original name: ukn |
| JANIS_CD | JANIS identifier. Original name: poolid |
| anon_BVD_CD | BVD identifier (anonymized) |
| anon_ENTTY_RIAD_CD | RIAD identifier (anonymized) |
| anon_LEI | LEI (anonymized) |
| anon_USTAN_CD | USTAN identifier (anonymized) |
| anehmernr | Borrower ID, anonymized (”Nehmernummer, anonymisiert”) |

## 2.2 Details of Variables

**AWMUS_CD:** AWMuS identifier. Anonymous. Original name: MLD_NR

| | |
|---|---|
| Notes | The Bundesbank database AWMuS is the masterdatabase for all foreign statistics related master and metadata in the Deutsche Bundesbank. Source of master data for the research datasets MiDi, SITS and SIFCT. Note that AWMUS_CD does not include the trailing check-digit included in the original MLD_NR. |
| Available from – to | 1979 – 2021 |
| Type of variable | String with up to 8 digits. |
| Anonymization | Bundesbank internal anonymous ID, may be used in research datasets. |

**BBK_H_CO_ID1:** JANIS identifier (anonymized)

| | |
|---|---|
| Notes | The Bundesbank database JANIS contains annual financial statements of German non-financial corporations. Successor to USTAN. Since this anonymization of the JANIS-ID is provided by the departement responsible for JANIS, the ”BBK_HM_CO_ID1” is not necessarily produced from the same data version, therefore there may not always be a ”BBK_HM_CO_ID1”-value for every ”USTAN_PLUS_CD”-value in the JANIS-Data available to the RDSC. |
| Available from – to | 1997 – 2019 |
| Type of variable | String with 10 digits, corresponds to the anonymized POOLID. |
| Anonymization | anonymized USTANPLUS_CD (anonymization is done by the data providing Bundesbank Department) |

**BVD_CD:** Bureau van Dijk identifier. Original name: bvdid

| | |
|---|---|
| Notes | The Bureau van Dik identifier is included in the data by the dataprovider Burau van Dijk. A part of the entries with this ID stems from BvD-Master data acquired directly from BvD. Since the BvD-ID is (for German companies) derived from the ID by the German data provider "Creditreform" (by adding the prefix "DE" to the ID-value), it can be complemented by master datasets from Creditreform. In the case of the data used by the RDSC, data from the "Mannheimer Unternehmenspanel" (MUP), from the Zentrum für Europäische Wirtschaftsforschung (ZEW), which is derived from Creditreform data, is used to complement the available BvD-Data. For this purpose, the Creditreform ID is transformed into the BvD-ID by adding the prefix "DE" to the ID values. In consequence, in the data that enters our record linkage, entries that stem from BvD as well as entries that stem from the MUP are identified by this ID. |
| Available from – to | 2004 – 2021 |
| Type of variable | String with up to 20 characters. |
| Anonymization | ID may not be used in anonymous research datasets. For an anonymized version of this ID see anon_BVD_CD |

**DE_BAKISN_CD:** Borrower ID ("Nehmernummer")

| | |
|---|---|
| Notes | The Bundesbank database "BAKIS-M" holds bank supervision reference data on borrowers. It contains 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. BAKIS-M is the source of master data for analytical and research datasets generated from BAKIS-M. |
| Available from – to | 2002 – 2018 |
| Type of variable | String of 7 or 8 digits, format as defined by the RIAD data standard. |
| Anonymization | ID may not be used in anonymous research datasets. There is no anonymized standard version of this ID. |

**DE_DESTATIS_CD_STBL:** Destatis business register ID. Original name: WE_ID_ALT

| | |
|---|---|
| Notes | The official business register ("URS") of the German Statistical Office ("DESTATIS") uses this ID. |
| Available from – to | 2012 – 2019 |
| Type of variable | String of 9 digits as defined by DESTATIS. |
| Anonymization | ID may not be used in anonymous research datasets. There is no anonymized version of this ID. |

**ENTTY_RIAD_CD:** RIAD identifier.

| | |
|---|---|
| Notes | The Bundesbank database "BBk-RIAD" (RIAD: "Register of Institutions and Affiliates Database") is the Bundesbanks' master database for AnaCredit (german part). |
| Available from – to | 2018 – 2021 |
| Type of variable | String with up to 50 characters, format as defined by the RIAD data standard. |
| Anonymization | ID may not be used in anonymous research datasets. There is no anonymized version of this ID yet, since there is no research dataset for AnaCredit yet. |

**LEI:** Legal entity identifier

| | |
|---|---|
| Notes | The LEI-data from the Global Legal Entity Identifier Foundation issues and maintains this identifier. |
| Available from – to | 2018 – 2021 |
| Type of variable | String with up to 20 characters as defined by the Local Operating Units of the Global Legal Entity Identifier System (GLEIS). |
| Anonymization | ID may not be used in anonymous research datasets. For an anonymized version of this ID see anon_LEI. |

**USTAN_CD:** USTAN identifier. Original name: ukn

| | |
|---|---|
| Notes | The Bundesbank database USTAN is a repository that includes master data on companies that have been reported to the Deutsche Bundesbank in the context of its refinancing operations and later for credit assessment purposes. Apart from master data, it contains HGB and IFRS annual financial statements for companies, insolvency data, data reported for the credit register and rating information. Source of master data for USTAN. For our record linkage, the master data in USTAN, which is mostly limited to recent years, is complemented by historical (historized) master data from the original source databases of USTAN ("Jalys" and "Cops"). |
| Available from – to | 1987 – 2018 |
| Type of variable | String with up to 8 digits. |
| Anonymization | ID may not be used in anonymous research datasets. For an anonymized version of this ID see anon_USTAN_CD. |

**JANIS_CD:** JANIS identifier. Original name: poolid

| | |
|---|---|
| Notes | The Bundesbank database JANIS contains annual financial statements of German non-financial corporations, including public data sources. Successor to USTAN. |
| Available from – to | 1997 – 2019 |
| Type of variable | String with 10 digits, corresponds to the anonymized POOLID. |
| Anonymization | ID may not be used in anonymous research datasets. For an anonymized version of this ID see BBK_H_CO_ID1. |

**anon_BVD_CD:** BVD identifier (anonymized)

| | |
|---|---|
| Notes | Anonymized version of BVD_CD. |
| Available from – to | 2004 – 2021 |
| Type of variable | String with 64 characters. |
| Anonymization | anonymized (SHA256 hash based on BVD_CD). |

**anon_ENTTY_RIAD_CD:** RIAD identifier (anonymized)

| | |
|---|---|
| Notes | Anonymized version of ENTTY_RIAD_CD |
| Available from – to | 2018 – 2021 |
| Type of variable | String with 64 characters. |
| Anonymization | anonymized (SHA256 hash based on ENTTY_RIAD_CD). |

**anon_LEI:** LEI (anonymized)

| | |
|---|---|
| Notes | Anonymized version of LEI. |
| Available from – to | 2018 – 2021 |
| Type of variable | String with 64 characters. |
| Anonymization | anonymized (SHA256 hash based on LEI). |

**anon_USTAN_CD:** USTAN identifier (anonymized)

| | |
|---|---|
| Notes | Anonymized version of USTAN_CD. |
| Available from – to | 1987 – 2018 |
| Type of variable | String with 64 characters. |
| Anonymization | anonymized (SHA256 hash based on USTAN_CD). |

**anehmernr:** Borrower ID, anonymized (”Nehmernummer, anonymisiert”)

| | |
|---|---|
| Notes | Anonymized version of DE_BAKISN_CD. |
| Available from – to | 2002 – 2018 |
| Type of variable | String with 8 characters. |
| Anonymization | Project specific anonymization procedure. |

## References

Bundesamt für Sicherheit in der Informationstechnik. (2020). Kryptographische Verfahren: Empfehlungen und Schlüsselängen. BSI – Technische Richtlinie, *BSI TR-02102-1, 2020-01*. Bonn, Germany: Federal Institute for Information Security.

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.

Gabor-Toth, E., & Schild, C.-J. (2021). Understanding Overlaps between Different Company Data. Technical Report 2021-06, Research Data and Service Centre, Deutsche Bundesbank.