// NO.26-056 | 11/2026

DISCUSSION PAPER

// ENRICO DE MONTE, LION HOLSTE, JAVIER MIRANDA,
VERENA PLÜMPE, AND JIAMIN WANG

From Legal Units to Firms: A Hybrid Approach to Ownership-Based and Name-Based Firm Identification

Leibniz Association

ZEW

From Legal Units to Firms: A Hybrid Approach to Ownership-Based and Name-Based Firm Identification*

Enrico De Monte[1], Lion Holste[1], Javier Miranda[2], Verena Plümpe[†2], and Jiamin Wang[2]

[1]Leibniz Centre for European Economic Research (ZEW)

[2]Halle Institute for Economic Research (IWH)

November 17, 2026

Abstract

Business data typically record establishments or legal units, while economic decisions often span multiple legal entities under common control. We develop a framework for the Mannheim Enterprise Panel that constructs time-consistent identifiers from ownership networks and business-name matching. The two approaches produce separate, complementary outputs, distinguishing common control from name similarity. Linking the ownership-based identifier to Germany's Establishment History Panel reveals the economic importance of enterprise boundaries: identified enterprises comprising multiple legal units account for more than 30% of total employment recorded in the panel by the end of the sample. An application decomposes establishment-level job creation and destruction by enterprise structure, revealing the organizational context of job flows while preserving conventional aggregate rates. The resulting crosswalks connect establishments, legal units, and enterprises, enabling longitudinal research on firm size, organization, and dynamics beyond legal-unit boundaries.

Keywords: firm identification; record linkage; ownership networks; enterprise boundaries; business registers; longitudinal business data.

JEL classification: C55; C81; D22; L22.

*This research is part of a joint IWH–ZEW project funded by the Leibniz Association through its Leibniz Collaborative Excellence programme (K525/2023). We thank Thorsten Doherr (ZEW) for his valuable contributions to the construction of the MUP–Ownership Panel and all project members for their contributions to the development of this research through workshops and discussions.

†Authors are listed in alphabetical order. Corresponding author: Verena Plümpe, vpe@iwh-halle.de.

1 Introduction

Firm disambiguation constitutes a fundamental challenge in the development of business databases derived from transaction-level reports on legal units. From an economic standpoint, identifying the actual unit of control – the firm or enterprise where business decisions are made – is essential to analyze business performance accurately. Without consistent longitudinal identifiers, it is impossible to distinguish between establishments, legal units, firms, or enterprises, thereby impeding coherent analysis of firm-level outcomes and dynamics.[1] Notably, neither the commercial business register (Handelsregister) nor the company register (Unternehmensregister) adequately capture firm structures. Moreover, changes in identifiers due to mergers, acquisitions, or technical reasons such as duplicate assignments by credit rating agencies further complicate the task.

Existing linkage methods predominantly rely on name-based matching and more rarely on ownership-based algorithms, both of which exhibit well-recognized limitations. Name-based approaches tend to under-detect affiliations where subsidiaries operate under different trade names, while ownership-based strategies face difficulties in addressing restructuring events, especially around smaller units where coverage is more limited. Recent methodological contributions apply probabilistic and machine learning techniques focusing separately on entities' names and addresses (Doherr 2023) or ownership structures (Reich 2024), yet resulting firm identifiers often remain fragmented or incomplete.[2] Ascione, Sterzi, and Vezzulli (2025) introduce beyond standard structure of NLP – including parsing, matching and a filtering stage – with the so-called *Terrorizer algorithm* a knowledge augmentation phase. Thereby, an additional external source of data is added to the data before the usual text-based processing, to verify that firm names belong to firms from the same enterprise. However, adding such knowledge augmentation to a longitudinal business database retrospectively might introduce a bias over time, as web-scraping is likely more effective for recent years.[3]

Regarding the German landscape of legal units and enterprises, an important recent contribution is provided by Diegmann et al. (2026), who build on the algorithms of Doherr (2023) in the form of a name and address-based record linkage between legal units from the IAB and enterprise data from the Mannheim Enterprise Panel (MUP). Nevertheless, the linked units in MUP still fall short of delivering consistent firm-level disambiguation (De Monte et al. 2026, forthcoming). More broadly, existing methodological approaches often rely on name-based

[1] These terms are defined in Section 2.

[2] Reich (2024) propose a machine learning and natural language processing approach to link German survey data with external financial data from the Orbis database, thereby improving match quality and offering detailed documentation.

[3] Arora, Belenzon, and Sheer (2021) used a bespoke combined database including data from ORBIS, SDC Platinum, and WRDS to match corporate subsidiaries to Compustat "ultimate owner" firms in the same spirit as the *Terrorizer* algorithm to their analysis of patents and scientific works.

matching since ownership-based network information is limited in US administrative data.[4] Ownership-based approaches use control links to construct ownership structures or networks, for example to study multinational firms, firms with subsidiaries, or industry concentration.[5] Despite these advances, which allow researchers to connect high-quality administrative data on legal units with enterprise data, establishing consistent longitudinal firm identifiers remains challenging, primarily due to identifier changes, corporate restructuring, and data-quality issues. Reliable linkage is indispensable for empirical research on firm dynamics, labor markets, and productivity, and aligns with statistical standards recently exemplified by the EU enterprise concept implementation in German structural business statistics (Beck et al. 2020) starting in 2018, reflecting the long-standing recognition by the U.S. Census Bureau of the firm as the correct unit of economic analysis, as applied in the construction of the Longitudinal Business Database (Jarmin and Miranda 2002).

In response, this paper develops a hybrid disambiguation framework tailored to the MUP. It combines community detection on ownership networks with fuzzy matching of business names to produce two time-consistent and complementary outputs: an ownership-based enterprise identifier and a name-based identifier. The framework is designed to reduce the artificial fragmentation that arises when each legal unit is treated as a separate firm, while preserving the conceptual distinction between control-based and name-based links. We document the coverage, overlap, and concordance of these outputs and use the ownership-based enterprise identifier to show how measured job reallocation differs between single-legal-unit (SLU) and multi-legal-unit (MLU) enterprises.

Research on wage setting illustrates the importance of these organizational boundaries. Using the MUP–BHP link, Diegmann, Müller, and Schoefer (2026) find that the establishment-size wage gradient disappears across establishments within firms defined as legal units. Our ownership-based identifier extends the observable firm boundary to legal units under common control, allowing researchers to examine whether relationships documented within legal units also extend across affiliated legal units within the same enterprise.

Building on these methodological advances, the paper is structured as follows. Section 2 defines the statistical units and terminology used throughout the paper. Section 3 presents the data sources that enter the disambiguation process. Section 4 describes the two-pillar algorithmic framework, including the ownership-based and name-based components and the rules used to impose temporal consistency. Section 5 documents the empirical properties of the

[4] Name-based disambiguation is common in patent and firm-linkage applications. Ascione, Sterzi, and Vezzulli (2025) summarize applications in the patent literature; Bremer (2023) link firms in the Amadeus Financials and Subsidiaries databases to applicants in PATSTAT using a four-step algorithm. Some approaches combine name similarity with auxiliary information to improve match quality, such as board similarity and HCO information (Garcia-Bernardo and Takes 2018).

[5] Examples include Garcia-Bernardo et al. (2017) on multinational firms, Cooper et al. (2016) on firms with subsidiaries, and Bajgar et al. (2023) and Calligaris et al. (2024) on industry concentration. Many such studies rely on Orbis data; limitations of Orbis without further processing are documented by Bajgar et al. (2020) and Arndt (2023).

resulting identifiers, including their coverage, cluster-size distributions, overlap, and concordance. Section 6 illustrates the economic relevance of the ownership-based enterprise identifier using linked employment data, including a decomposition of job reallocation between SLU and MLU enterprises. Section 7 concludes.

2 Statistical Units and Firm Identification

We begin by clarifying the statistical units that enter the construction of firm identifiers. Administrative and commercial business data typically record activity at the level of establishments or legal units, whereas many economic questions require a unit that more closely reflects common control and coordinated decision-making.[6] This distinction is central to the statistical "unit problem" and to the implementation of the enterprise concept in business statistics (Beck et al. 2020; Eurostat 2024; Sturm 2018). Neither the commercial business register nor the German company register (Unternehmensregister) provides such a time-consistent firm identifier directly. We therefore construct identifiers from observed ownership, business-name, and activity information.

Conceptually, we follow the distinction between establishments and firms used by the U.S. Census Bureau and discussed by Haltiwanger, Jarmin, and Miranda (2013). A *local unit*, or establishment, is a physical location at which business activity takes place, whereas a firm comprises the establishments under common operational control.

In the IAB data, employment is observed for administrative establishment identifiers (*betnr*) in the Establishment History Panel (BHP). These reporting units are designed to represent establishments, but they need not correspond perfectly to a single physical location and may coincide with the legal unit for single-location businesses. We therefore use *BHP establishment* for the employment-reporting unit identified by *betnr*, and reserve *legal unit* for a formally registered entity identified by a Creditreform *crefo*. A legal unit may be linked to one or several BHP establishments. An *enterprise* is a legal unit or a group of legal units under common ownership or control. In small businesses, the BHP establishment, legal unit, firm, and enterprise may coincide. In more complex organizations, several BHP establishments may belong to one legal unit and several legal units may belong to one enterprise.

A useful distinction for the empirical application is between *single-legal-unit (SLU) enterprises* and *multi-legal-unit (MLU) enterprises*. An SLU enterprise contains exactly one legal unit in a given year, whereas an MLU enterprise contains more than one legal unit in that year. This classification is made at the enterprise-year level and is conceptually distinct from the number of establishments. An SLU enterprise may contain a single legal unit linked to one or

[6] Legal-unit records need not coincide with the economically relevant decision-making unit. They may be more aggregated than establishment-level activity when one legal unit operates multiple local units, or more disaggregated than the relevant enterprise when several legal units are under common control.

several BHP establishments; an MLU enterprise contains several legal units, each of which may itself be linked to one or several BHP establishments. A legal unit may therefore belong to an SLU enterprise in one year and to an MLU enterprise in another year if the enterprise structure changes.

The objective of our algorithm is to construct time-consistent identifiers that approximate economically meaningful decision-making units within this hierarchy. We define a firm, in this empirical sense, as a legal unit or cluster of legal units that is plausibly subject to common control and coordinated business decisions. This unit is the relevant object for studying firm-level outcomes because many decisions about investment, employment, innovation, and restructuring are made above the level of the individual legal unit.

Operationally, the construction combines several pieces of information. First, legal units are grouped when ownership relations indicate unified control and, where applicable, a common highest controlling unit (HCU). Second, business-name information is used to identify additional links across legal units that may not be captured by observed ownership relations.[7] Third, the MUP–LBD provides information on whether legal units are active, and its link to IAB establishment data allows us to relate legal units to observed employment activity (Diegmann et al. 2026). We refer to this linked data environment as the MUP–BHP. Finally, temporal continuity is imposed by comparing cluster composition across adjacent years using Jaccard similarity.

This definition is deliberately empirical. It does not impose a one-to-one correspondence between legal units, firms, and enterprises. Instead, it uses observed ownership, business-name, and activity information to construct identifiers suitable for longitudinal economic analysis. Temporal consistency matters because legal-unit clusters can change for reasons that should not necessarily imply the birth of a new firm identifier. For example, a highest controlling unit may change while most legal units in the cluster remain together, or parts of a group may be divested while the remaining units continue under common control. The longitudinal step is designed to preserve continuity in such cases while still allowing identifiers to change when the underlying firm structure changes. The resulting identifiers are intended to capture units at which business decisions are likely to be coordinated, while avoiding the artificial fragmentation that arises when each legal unit is treated as a separate firm.

3 Data

We rely on data provided by Creditreform, Germany's largest credit rating agency, which collects extensive information on businesses active in Germany. The raw data are provided to the Leibniz

[7] Name-based linkage and consolidation methods are also used in related settings, including company-data linkage and patent-assignee disambiguation; see, for example, Doherr (2023), Bremer (2023), and Ascione, Sterzi, and Vezzulli (2025).

Centre for European Economic Research (ZEW), where they are processed for research purposes into several data products.[8] To disentangle business structures in the MUP, we use three of these products: the MUP–Ownership Panel, the MUP Longitudinal Business Database (MUP–LBD), and the MUP–Business Name Panel.

3.1 MUP–Ownership panel

We first process Creditreform data on ownership and shareholders of observed legal business units, identified by *crefo* IDs, to derive the MUP–Ownership Panel. A central source for these ownership data is the commercial business register, where German business units are required to register and disclose shareholder information depending on their legal form. Limited liability companies (GmbH) and partnerships (OHG, KG) are generally required to provide this information. Public limited companies (AG) typically disclose founding owners and shareholders; publicly traded companies are required to report ownership information only if a majority shareholder exists. Registered businesses must update their ownership and shareholder structure whenever a relevant change occurs.[9]

Creditreform transmits these data to ZEW twice per year, in January and July. Each transmission contains updates to ownership or shareholder information, as well as newly observed business units.[10] We use these transmissions to construct an annual ownership panel by carrying forward the most recent ownership information for each *crefo* when no update is observed in the current year.[11] Updates received in the January transmission are assigned to the previous calendar year, and only the latest ownership information for a given *crefo* in a given year is retained.

Owners are listed under a so-called b_crefo ID, which refers either to a physical person or to another legal business unit. Corporate owners can be domestic or foreign. If the shareholder is a domestic company, it is typically also observed as a *crefo* in the MUP–LBD, the second dataset we use.

3.2 MUP–LBD

The MUP–LBD is an annual panel of legal business units, identified by *crefo* IDs, that are active in Germany during the period covered by our analysis (De Monte et al. 2026, forthcoming). It is constructed from the raw Creditreform data, which contain a much larger and increasing number of observed *crefo* units over time. The raw files do not directly correspond to the stock of active

[8] See De Monte et al. 2026, forthcoming for more information.

[9] This includes changes in the set of owners and changes in ownership shares.

[10] Coverage depends on observed changes in the source data and may differ by legal form because disclosure requirements vary across firm types.

[11] If Creditreform updates shareholder information for a *crefo* more than once within a year, we keep the update closest to June 30 to align the ownership data with the timing of the IAB data and to reduce noise.

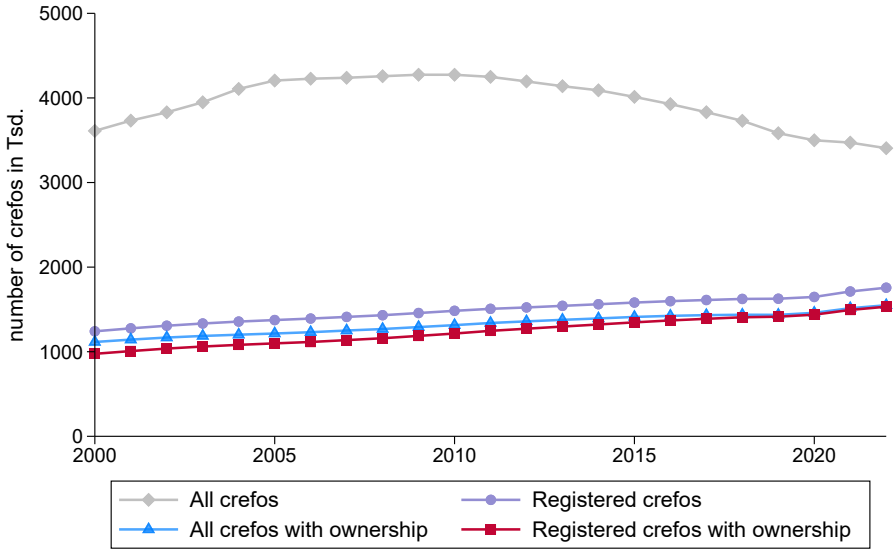


Figure 1: Coverage of Shareholder Information for Legal Units Observed in the MUP–LBD

Notes: The figure shows the total number of legal business units observed in the MUP–LBD, the subset registered in the commercial business register, and the subset for which shareholder information is available in the MUP–Ownership Panel. Ownership records are observed longitudinally; if no change is reported, the most recent ownership information is carried forward.

businesses in Germany because they may include foreign units, duplicate records, units that ceased activity before the sample period, and units for which deregistrations or exits are rarely observed.[12]

The MUP–LBD addresses these issues and restricts attention to businesses active in Germany. This yields about 8.7 million active legal business units over the full period, of which about 46% (approximately 4 million) are registered in the commercial business register. Overall, we observe 3.1 million legal business units with ownership information. A key advantage of the MUP–LBD is that it contains information on entry and exit dates, which we use in the disambiguation algorithm. In constructing the MUP–Ownership Panel, this allows us to condition ownership information on whether a unit is active in the MUP–LBD.[13]

Figure 1 depicts the corresponding time series. The total number of legal units observed in the MUP–LBD changes over time, increasing from about 3.6 million units in 2000 to approximately 4.2 million in 2009 and then declining to about 3.3 million in 2023. By contrast, the number of units registered in the commercial business register rises steadily, from about 1.2 million in

[12] Some observed units may have ceased activity before the sample period, while others may have registered once without subsequent activity updates. This may occur for foreign businesses or for units in neighboring regions also covered by Creditreform, such as Austria and Liechtenstein. In addition, Creditreform has a decentralized structure in which local branches collect information on business activity in their region, including some neighboring foreign regions. In some cases, different branches collect information on the same business under distinct *crefo* IDs. Known duplicates are flagged by Creditreform. See De Monte et al. (2026, forthcoming) for details.

[13] A *crefo* may appear only once in a Creditreform ownership update, with ownership information but without subsequent activity reports. The MUP–LBD is therefore needed to distinguish ownership records attached to active businesses from records attached to units no longer active in Germany.

2000 to 1.7 million in 2023. The number of units for which ownership information is available is slightly lower than the number of registered units.[14] Our ownership-based approach therefore relies mainly on registered businesses. The implications of missing ownership information for non-registered businesses and publicly traded corporations are discussed in Section 3.5.

3.3 MUP–Business Name Panel

The third data source is the MUP–Business Name Panel, which provides longitudinal information on business names recorded by Creditreform. These records associate legal business units with the names under which they are observed in the data. When a legal unit changes its name, the new name may appear as an additional record. Such changes can arise in the course of mergers, acquisitions, legal reorganizations, or changes in business activity.

The panel is constructed in the same spirit as the MUP–Ownership Panel. For each *crefo*, we create an annual name record. If no new name is observed, the most recent name is carried forward. Because deletions of name records are rare, the panel preserves observed historical name information and accumulates over time. The resulting data should therefore be interpreted as a longitudinal record of observed business-name information, rather than as a complete event history of legal name changes.

The MUP–Business Name Panel provides the main input for the name-based component of the disambiguation algorithm. It is particularly valuable for identifying links among legal units when ownership information is missing, incomplete, or insufficient to establish common control. At the same time, business names alone do not identify economic activity, firm exit, or ownership control. For this reason, name-based information is used in combination with the MUP–Ownership Panel and the MUP–LBD when constructing time-consistent firm identifiers.

3.4 Raw Data Coverage

Figure 2 documents the overlap between ownership and business-name information in the raw inputs to the disambiguation procedure. Panel A reports coverage in the full comparison sample of *crefos* used to assess raw data availability before imposing the commercial-register restriction. Panel B restricts the sample to *crefos* observed in the commercial business register. In both panels, each covered *crefo* is assigned, by year, to one of three mutually exclusive categories: covered only by ownership information, covered only by business-name information, or covered by both sources.

The figure should therefore be read as a coverage comparison, not as an annual stock of active legal units. Raw business-name records accumulate over time, and not all *crefos* observed in these raw name files are necessarily found in the MUP–LBD. Once a business name or ownership

[14] This close correspondence is expected because Creditreform primarily extracts ownership information from the commercial business register (Handelsregister).

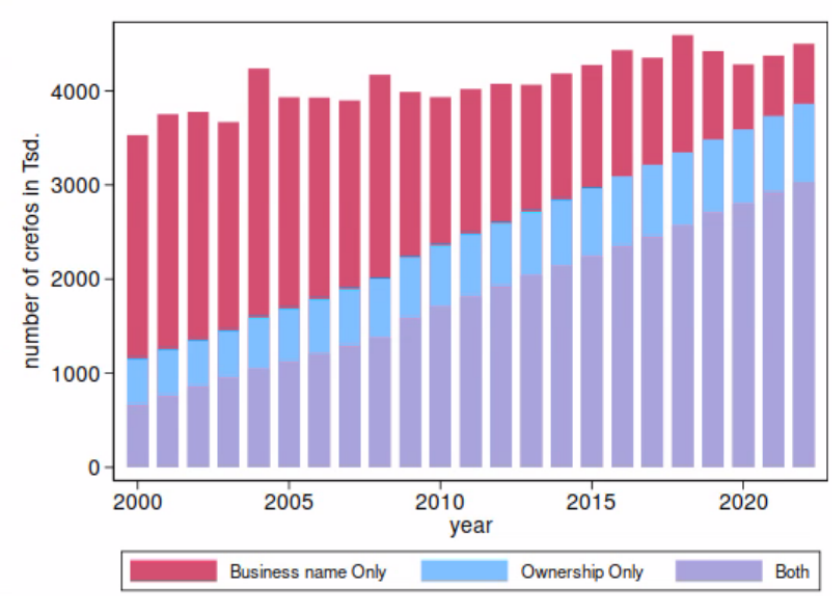
link is observed, it remains available for the coverage measure. The exception is a *crefo* that is observed in the MUP–LBD and has an exit event there: in that case, we exclude the unit from the coverage count in the exit year and all subsequent years. Consequently, the increase in coverage over time primarily reflects the arrival and retention of additional business-name and ownership information, together with this MUP–LBD exit restriction, rather than entry and exit dynamics in the full population of legal units.[15]

The central pattern is the growing overlap between ownership and business-name information. In the full comparison sample, business-name-only observations account for a large share of coverage at the beginning of the period, but the number of *crefos* observed in both sources rises steadily over time. The overlap is even more pronounced among commercial-register units, where the "both sources" category is already large in the early 2000s and becomes the dominant component of coverage thereafter. This common support is important for the design of the algorithm: for units observed in both data sources, ownership-based and name-based signals can be compared directly, allowing the algorithm to confirm consistent links and to flag cases in which similar names do not correspond to common control.

At the same time, neither source is nested in the other. In the full comparison sample, ownership-only coverage remains sizeable, indicating that ownership records provide an independent source of information for identifying firm structures rather than merely reproducing the information available from business names. Name-only coverage also remains relevant, especially outside the commercial-register sample, because it identifies units that can enter the name-based pillar even when ownership information is unavailable.

[15] The algorithmic implications of this changing information set are addressed below. Clustering is performed on a cross-sectional, year-by-year basis and then harmonized longitudinally across years.

(a) Full comparison sample



(b) Commercial business register

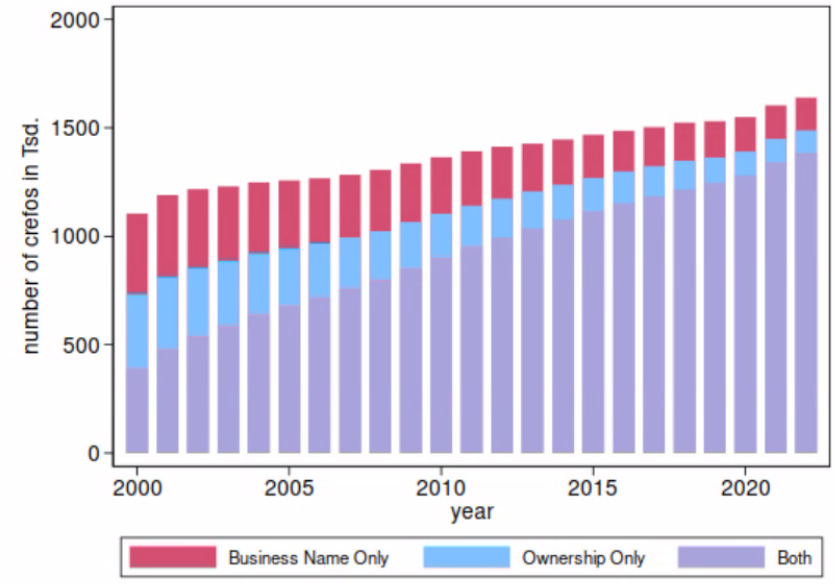


Figure 2: Coverage of Ownership and Business-Name Information

Notes: Panel A shows coverage of ownership and business-name information in the full comparison sample. Panel B restricts the sample to *crefos* observed in the commercial business register. Bars count unique covered *crefos* in the respective year and distinguish three mutually exclusive categories: *crefos* covered only by ownership information, only by business-name information, and by both sources. Ownership information refers to records from the MUP–Ownership Panel. Business-name information refers to records from the MUP–Business Name Panel. Raw business-name and ownership inputs are retained once observed. If a *crefo* is observed in the MUP–LBD and has an exit event there, it is excluded from the exit year onward. The figure should therefore not be interpreted as an annual stock of active firms.

Overall, Figure 2 provides the empirical motivation for combining the two data sources. The large common support between ownership and business-name information enables internal

cross-checks between the two algorithmic pillars. The non-overlapping components, in turn, show that each source adds distinct coverage. A purely ownership-based or purely name-based procedure would therefore discard useful information available in the other source.

In the commercial-register sample, incremental coverage from either source alone is more limited, reflecting the high overlap between ownership and name information for registered legal units. Nevertheless, the presence of ownership-only and name-only observations in this restricted sample shows that the two inputs remain complementary even among units for which administrative coverage is strongest. Ownership-only observations contribute control information that would otherwise be missing from name-based matching, while name-only observations allow the algorithm to exploit naming information when shareholder records are absent or incomplete.

3.5 Data Limitations

The ownership data provide broad coverage of German registered legal units, but two limitations are particularly relevant for the construction of ownership-based identifiers. The first concerns selective reporting requirements across legal forms. The second concerns the domestic scope of the ownership records.

Many legal units without ownership information are sole proprietorships (Einzelunternehmen) or small businesses (Kleingewerbe), which are generally not required to register in the commercial business register and typically have less complex ownership structures. Missing ownership information for these units is therefore less likely to omit economically meaningful control links. For publicly traded corporations with dispersed shareholdings, the absence of a reported controlling shareholder is not itself a limitation of the approach: if no entity exercises control, the unit should not be assigned to a higher controlling unit. The relevant limitation arises when the data do not fully capture transitions between majority ownership and dispersed ownership. Units without an observed controlling owner can nevertheless enter firm clusters as shareholders if they themselves act as majority owners of other legal units.

This reporting rule can also generate measurement error when ownership structures change over time. Consider a publicly traded company A that is majority-owned by company B in year $t-1$, with this ownership relation recorded in the commercial business register. If B sells its shares in A in year t to a dispersed set of investors, A is no longer required to report a controlling shareholder. Because this divestment may not be recorded as an explicit ownership update, the register may continue to list B as the majority owner. In this case, the data would erroneously maintain the link between A and B, leading to an overstatement of the corresponding ownership-based firm cluster.

A second limitation is that the ownership data primarily cover German-registered legal units. Foreign controlling entities may appear as shareholders, but their full ownership structures are

not observed.[16] Acquisitions by foreign entities are captured when they involve German legal units, but ownership chains terminate at the highest observable shareholder in the available data rather than at the ultimate global parent. This is not central for our purposes, since our analysis does not require identifying ultimate global ownership structures.

4 Firm Disambiguation Algorithm

This section describes the algorithm used to construct time-consistent identifiers from Creditreform data on ownership relations and business names. The basic unit of observation is the *crefo*, which identifies a legal business unit. The objective is to group *crefos* into economically meaningful units that better approximate common control and coordinated decision-making, and to assign identifiers that remain stable over time unless the underlying structure changes substantively.

The algorithm combines two complementary sources of information. The first is ownership information, which identifies legal units under common control. The second is business-name information, which identifies links among legal units that may not be captured by observed ownership relations. Each pillar first produces cross-sectional clusters. These clusters are then linked over time using similarity-based rules to construct longitudinally consistent identifiers.

The remainder of this section proceeds as follows. Section 4.1 introduces the two-pillar framework. Section 4.2 describes the Jaccard similarity index, which is used to impose temporal consistency. Section 4.3 presents ownership-based clustering from shareholder data. Section 4.4 describes the name-based clustering procedure. Section 4.5 discusses how the two sources should be interpreted jointly.

4.1 Two-Pillar Algorithmic Framework

This subsection outlines the conceptual structure of the disambiguation algorithm, summarized in Figure 3. The algorithm takes as inputs two types of information at the *crefo* level, provided by Creditreform: shareholder relations and business names. These inputs are processed separately in two algorithmic pillars. The resulting identifiers remain available separately and can be compared and interpreted jointly depending on the application.

The first pillar uses shareholder information to construct ownership-based clusters. We define control as majority ownership: a legal unit is treated as controlling another legal unit if it owns more than 50% of its shares, either directly or indirectly. In each cross-section, we identify ownership clusters by grouping all *crefos* that are ultimately controlled by the same entity. This step is implemented as an iterative, multi-layer procedure that captures complex ownership structures in the shareholder network. After each layer, the algorithm aggregates the ownership

[16] Foreign owners occasionally appear with a basic identifier but without further information.

shares held by legal units within a cluster. If the aggregated ownership share of cluster members exceeds 50% in another legal unit, that unit is added to the cluster in the subsequent layer.

After assigning *crefos* to ownership clusters in each cross-section, we impose temporal consistency. For each pair of adjacent years, ownership clusters are compared using the Jaccard similarity index and information on the highest controlling unit (HCU), as described below. These criteria are combined to reduce both false links between distinct firms and artificial breaks in continuing firm identifiers. High Jaccard similarity indicates continuity in cluster composition, while HCU information helps preserve continuity when the cluster remains under the same controlling entity despite minor changes in composition. The intermediate output of this pillar is a time-consistent ownership-based identifier, referred to as the *Ownership ID*.

The second pillar uses business names from the MUP–Business Name Panel to construct name-based clusters. Before matching, business names are standardized through parsing and harmonization routines and then tokenized, as described in Section 4.4.1. The harmonized names are matched using RECLINK2 in Stata.[17] To improve match quality, the algorithm excludes names that are too short, too generic, or dominated by highly frequent tokens, since such names can generate spuriously high similarity scores. Entities that are unlikely to belong to larger business structures, such as natural persons, sole proprietors, and certain associations, are also excluded from fuzzy matching and handled through exact string matching. These exclusion rules are described in Section 4.4.2.

The remaining business names are matched using fuzzy name matching, and the resulting pairwise links are grouped into annual name-based clusters, as detailed in Section 4.4.3. As in the ownership pillar, these cross-sectional clusters are linked across adjacent years using Jaccard similarity. If two name-based clusters exhibit sufficient overlap over time, they are assigned a common time-consistent identifier. The intermediate output of this pillar is the *Name ID*.

The next step compares the ownership-based and name-based identifiers. Name-based matches can be evaluated using ownership information whenever ownership data are available. If two name-matched *crefos* belong to the same ownership cluster, the match is confirmed by observed control relations. If they belong to different ownership clusters, the match is treated as evidence of name similarity rather than common control. This distinction is particularly useful for identifying franchise-type structures, in which legally independent firms may operate under identical or highly similar names.

[17] For a comparison of record-linkage tools in Stata and further details on RECLINK2, see Wasi and Flaaen (2015).

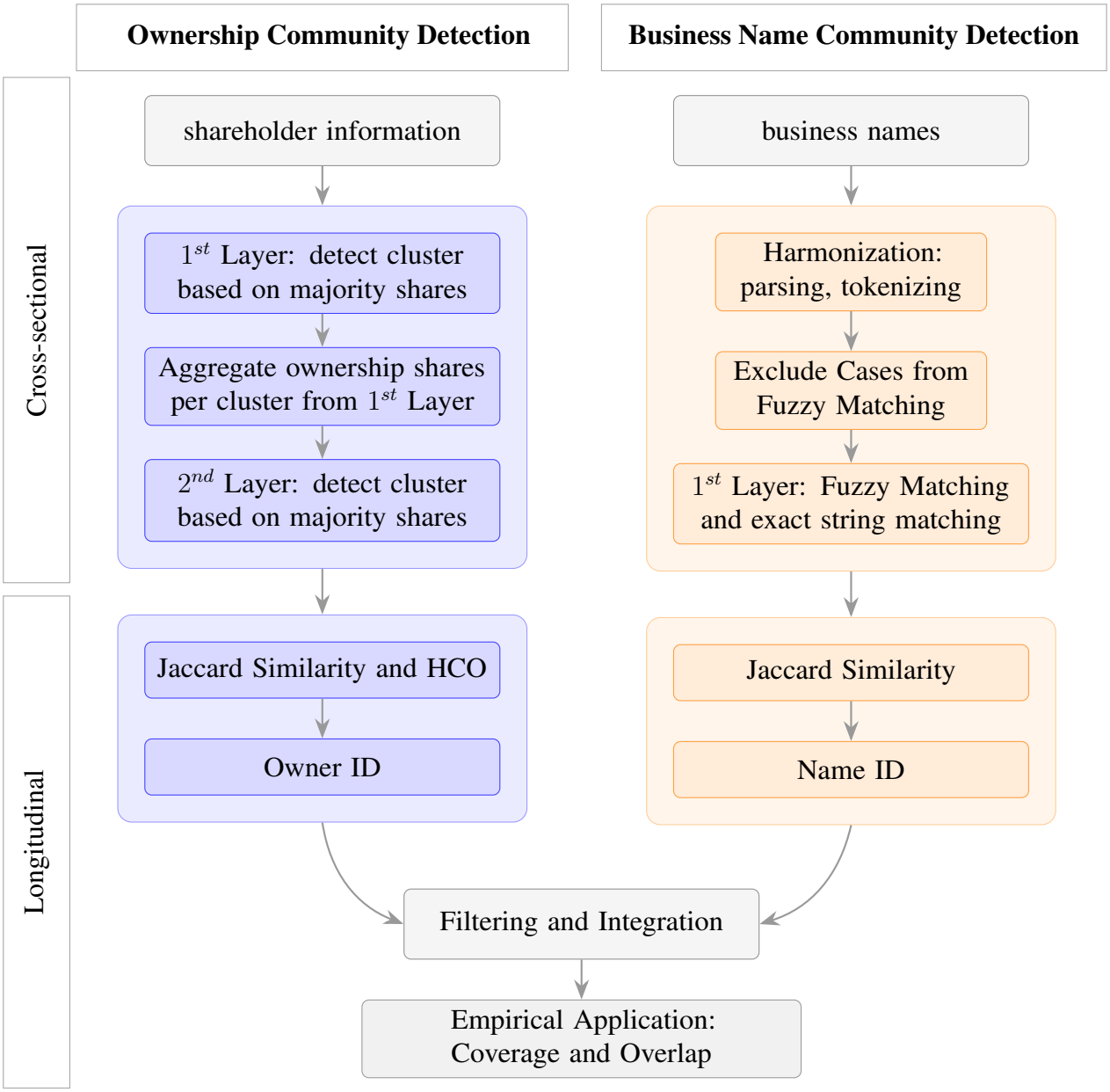


Figure 3: Two-Pillar Disambiguation Framework

Notes: The figure summarizes the construction of the two time-consistent identifiers. Both pillars take the *crefo* legal unit as their basic observation and first form clusters separately in each year. The ownership-based pillar uses shareholder links and majority-control relations to construct annual ownership clusters, which are linked over time using Jaccard similarity and continuity in the highest controlling unit (HCU), yielding the Ownership ID. The name-based pillar standardizes business names, constructs exact and fuzzy pairwise links, groups connected legal units into annual name clusters, and links these clusters over time using Jaccard similarity, yielding the Name ID. The two identifiers remain separate outputs and are compared where both sources are available; agreement provides corroborating evidence on enterprise boundaries, while disagreement is retained as economically informative rather than resolved through mechanical merging.

The joint interpretation of the two pillars exploits their complementarities. Ownership information identifies legal units under common control even when their names differ. Name matching identifies additional links where ownership information is missing or incomplete. Using both sources therefore improves coverage while reducing the risk of false positives from name similarity alone.

Finally, the intermediate *Ownership ID* and *Name ID* remain available as separate outputs. This separation is useful because ownership and business names capture different aspects of firm structure. The two identifiers can nevertheless be compared, and the section below provides guidelines for interpreting agreement and disagreement between them. The empirical application, including coverage, overlap between the two pillars, and links to the MUP–BHP, is reported separately below.

4.2 Temporal Consistency via Jaccard Similarity

The Jaccard similarity index is the central tool used to impose temporal consistency on the cross-sectional clusters generated by both pillars of the algorithm. Whether clusters originate from ownership relations or from business-name matching, the problem is the same: cross-sectional clustering produces a set of *crefos* in each year, but longitudinal analysis requires an identifier that can be followed over time. The Jaccard index provides a transparent measure of overlap in cluster composition across adjacent years.

For two clusters A and B, the Jaccard similarity is defined as

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

where A and B denote the sets of *crefos* assigned to two clusters in adjacent years. The index ranges from zero to one. A value of zero indicates no overlap in cluster composition, while a value of one indicates identical composition.

Starting from the initial cross-section in 2000, we compare each cluster in year t with candidate clusters in year $t + 1$ and identify the closest match according to the Jaccard index. If the similarity exceeds the relevant threshold, the cluster retains its identifier in year $t + 1$. If the similarity falls below the threshold, a new identifier is assigned. This procedure is then applied recursively to all adjacent year pairs.

The purpose of this rule is to distinguish minor changes in cluster composition from substantive reorganizations. Small additions or exits of individual *crefos* should not necessarily break the longitudinal identity of a firm cluster. By contrast, large changes in composition should lead to a new identifier. The Jaccard index provides a transparent and reproducible criterion for making this distinction.

Figure 4 illustrates the procedure. In 2000, *crefos* 1–6 form cluster A. In 2001, *crefo* 6 separates from the cluster, but the remaining units continue to overlap sufficiently with the

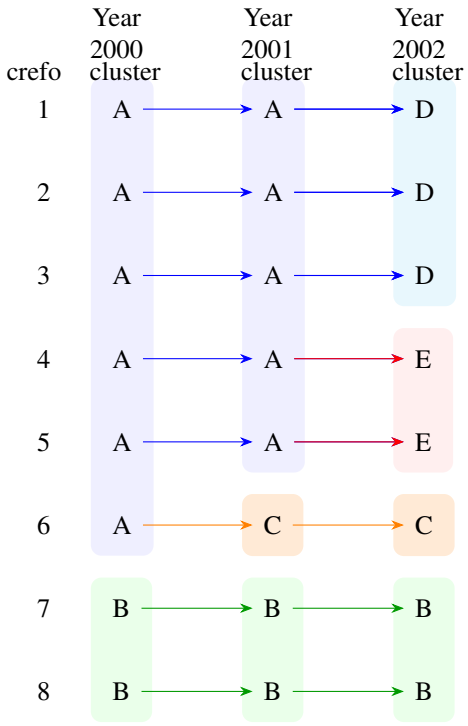


Figure 4: Jaccard Similarity: Application Example

Notes: The figure illustrates the assignment of *crefos* to time-consistent cluster identifiers. Clusters in adjacent years retain the same identifier if their Jaccard similarity exceeds the relevant threshold. In the example, *crefos* 1–6 form cluster A in 2000. In 2001, *crefo* 6 leaves the cluster, but the overlap remains sufficiently high for the continuing units to retain identifier A. In 2002, *crefos* 4 and 5 also leave, reducing the overlap below the threshold and leading to new identifiers. *Crefos* 7 and 8 remain in the same cluster throughout and retain identifier B.

previous-year cluster, so the identifier A is retained. By 2002, additional units leave the cluster and the overlap falls below the threshold, leading to the assignment of new identifiers. In contrast, *crefos* 7 and 8 remain jointly assigned throughout the period and retain the same cluster identifier.

Although the Jaccard routine is common to both pillars, the threshold and complementary decision rules differ across applications. Ownership-based clusters use a stricter threshold and are further refined using information on the highest controlling unit. Name-based clusters use the same similarity logic but rely only on overlap in name-based cluster composition. The next sections describe these pillar-specific implementations in detail.

4.3 Ownership-Based Clustering from Shareholder Data

This subsection describes the ownership-based pillar of the disambiguation algorithm. The objective is to identify groups of legal units that are subject to common control. We define control through majority ownership: a *crefo* is treated as controlling another *crefo* if it holds more than 50% of its shares, either directly or indirectly. The highest controlling unit (HCU)

provides the hierarchical anchor of the resulting ownership structure.[18]

The procedure has two steps. First, we identify ownership clusters separately in each cross-section by tracing direct and indirect majority-control relations. Second, we link these cross-sectional clusters over time using Jaccard similarity and HCU information. The cross-sectional step establishes which legal units are under common control in a given year; the longitudinal step determines whether such clusters represent the same firm over time.

4.3.1 Cross-Sectional Identification

The cross-sectional algorithm starts from direct majority-share relations. In the first layer, all legal units connected through direct or indirect majority ownership are assigned to the same cluster. This step captures control chains in which one legal unit owns more than 50% of another unit, possibly through intermediate majority-owned entities.

Figure 5 illustrates the logic. Solid arrows denote majority-share links and therefore imply control. For example, *crefo* 99999 directly controls *crefos* 10001 and 10002 and indirectly controls *crefo* 20001 through *crefo* 10001. These units are therefore assigned to the same first-layer ownership cluster.

In a second layer, the algorithm aggregates the ownership shares held by all members of each first-layer cluster. This step captures cases of joint control within a commonly controlled group. If the combined ownership share of cluster members in another legal unit exceeds 50%, that legal unit is added to the cluster. In Figure 5, orange dashed arrows indicate minority stakes that jointly imply control once aggregated at the cluster level. Gray dashed arrows denote minority stakes that do not imply control, even after aggregation.

The example also illustrates singleton clusters. *Crefo* 30001 is linked to other units only through minority ownership stakes that do not imply control and therefore remains a single-unit cluster. In the example, the 14 legal units are grouped into four ownership clusters based on majority-control relations. Appendix Table A.1 reports the same example in tabular form and shows how cluster identifiers are assigned across algorithmic layers.

The cross-sectional procedure therefore produces two complementary objects: the set of legal units under common control and the corresponding highest controlling unit. Both are used in the longitudinal assignment of ownership-based identifiers.

4.3.2 Longitudinal Identification: Jaccard Similarity and HCU

Following the cross-sectional assignment of *crefos* to ownership-based clusters, we construct time-consistent ownership identifiers. The purpose of this step is to distinguish genuine changes in firm structure from small or technical changes in observed cluster composition. A longitudinal

[18] The HCU is conceptually related to the global ultimate owner (GUO) used in other studies and databases such as Orbis. In our setting, however, it refers to the highest controlling unit identifiable within the available ownership data, not necessarily to the ultimate global parent.

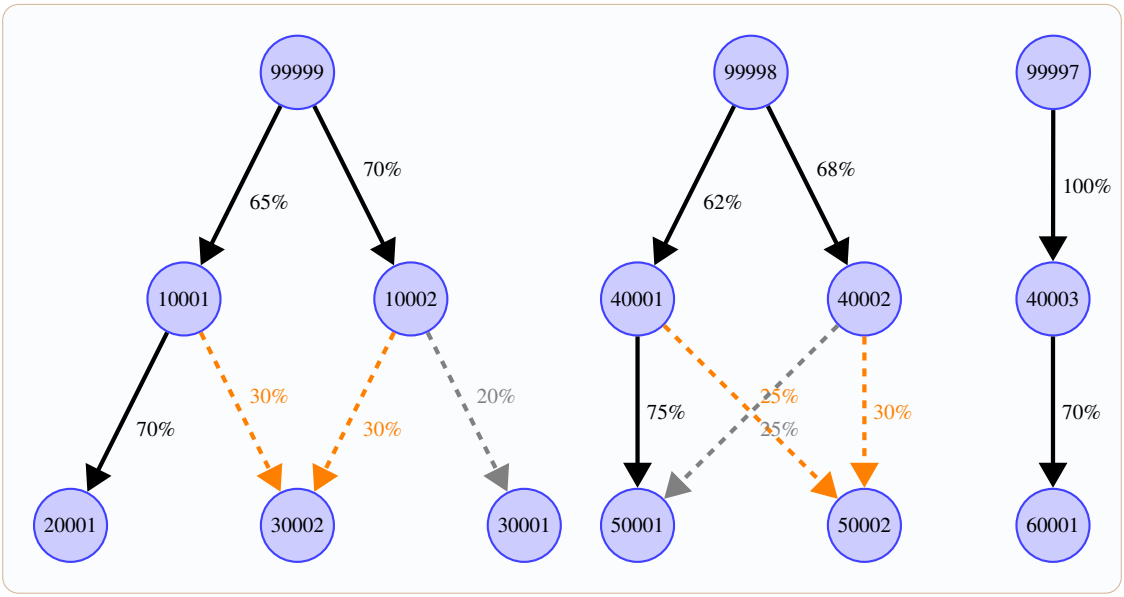


Figure 5: Clustering Based on Majority Ownership

Notes: The figure illustrates ownership and control relations among legal units. Solid arrows denote majority-share relations above 50% and therefore imply control. Legal units connected by solid arrows are grouped in the first layer of the algorithm. Orange dashed arrows denote minority stakes that jointly imply control after aggregating ownership shares within a first-layer cluster. Gray dashed arrows denote minority stakes that do not imply control, even after aggregation. Appendix Table A.1 reports the same example in tabular form.

identifier should remain stable when the underlying firm continues, but should change when the ownership structure changes sufficiently to indicate a new firm entity. The procedure therefore faces a standard trade-off between Type I error, falsely linking distinct firms, and Type II error, falsely splitting continuing firms into separate identifiers.

The baseline linking rule uses the Jaccard similarity between ownership clusters in adjacent years. For ownership-based clusters, we use a threshold of 0.85. Thus, a cluster in year t retains its ownership-based identifier in year $t + 1$ if its most similar counterpart exceeds this threshold. The threshold is intentionally strict because ownership-based clusters are intended to capture common control, and false links between distinct ownership structures would be particularly problematic for downstream firm-level analysis.

We refine this baseline rule using information on the highest controlling unit. The HCU provides an additional hierarchical signal that is especially useful when cluster composition changes only modestly or when the formal controlling unit changes within an otherwise stable ownership structure. Combining Jaccard similarity and HCU information yields four cases.

First, when Jaccard similarity exceeds 0.85 and the HCU is unchanged, the cluster is treated as continuing and retains its ownership-based identifier. In this case, both the composition-based and hierarchy-based signals indicate continuity.

Second, when Jaccard similarity exceeds 0.85 but the HCU changes, we prioritize compositional stability and retain the identifier. This rule avoids artificial breaks in cases where a

formal change in the controlling unit occurs within an otherwise persistent ownership structure. It therefore reduces Type II error by limiting the risk of spuriously splitting continuing firms into separate identifiers.

Third, when Jaccard similarity falls below 0.85 but the HCU remains unchanged, we preserve the identifier if no alternative cluster exceeds the similarity threshold with respect to the focal cluster. This rule is designed to avoid unnecessary breaks, particularly in smaller clusters where the entry or exit of a single *crefo* can mechanically generate a large change in the Jaccard index. In such cases, the unchanged HCU provides evidence of continuing control even though the composition-based threshold is not met. This refinement again limits Type II error.

Fourth, when both signals indicate discontinuity—that is, when Jaccard similarity is below the threshold and the HCU changes—we assign a new ownership-based identifier. This case is treated as a structural break in the ownership-based firm cluster, reducing the risk of Type I error from incorrectly linking distinct ownership structures.

This combined rule balances Type I and Type II errors in the longitudinal assignment of ownership-based firm identifiers. The strict Jaccard threshold limits Type I error: falsely assigning the same identifier to distinct firm structures. The HCU refinement limits Type II error: falsely assigning different identifiers to a continuing firm because of minor changes in observed legal-unit composition or changes in the formal position of the controlling unit. The resulting ownership-based identifier is therefore designed to be conservative in linking distinct structures while preserving continuity for persistent ownership groups. Appendix A.1 provides further details.

4.4 Name-Based Clustering from Business Names

This subsection describes the name-based pillar of the disambiguation algorithm. The objective is to identify links among legal units that are not captured by ownership information alone. Such cases arise when shareholder records are missing or incomplete, when ownership information does not identify a controlling entity, or when legal units are connected through naming conventions rather than observed majority ownership. Name-based clustering is therefore complementary to ownership-based clustering.

The input to this pillar is the MUP–Business Name Panel. We first harmonize raw business names by standardizing spelling, extracting legal-form and location information, and flagging tokens that may be uninformative for matching. We then exclude names that are too short, too generic, or otherwise unsuitable for fuzzy matching. These observations are handled separately through exact string matching. The remaining names enter a fuzzy matching procedure using RECLINK2. Pairwise name matches above the similarity threshold are grouped into annual name-based clusters. Finally, these cross-sectional clusters are linked across years using Jaccard similarity to construct time-consistent name-based identifiers.

4.4.1 Harmonization of Business Names

We harmonize raw business names before applying name-based matching. The purpose of this step is to reduce mechanical variation in spelling and formatting while preserving economically informative name components.

We first apply basic parsing routines to capitalized names. German special characters are replaced by their closest letter combinations, non-letter characters are removed, and leading, trailing, and duplicate spaces are eliminated. These steps standardize the raw strings and reduce noise from purely typographical differences.

We then identify legal-form expressions and extract them into separate tokens. Legal forms may appear as abbreviations, full expressions, or compound terms, and may occur either at the end or in the middle of a business name. Examples include the full designation "GESELLSCHAFT MIT BESCHRAENKTER HAFTUNG", the abbreviation "GMBH", and compound expressions such as "VERWALTUNGSGESELLSCHAFT MBH" or "VERWALTUNGS GMBH". Structural terms such as "VERWALTUNG" and "BETRIEBS" are retained as separate tokens because they may carry information about the role of a legal unit within a firm group.

Location information contained in business names is identified using a dictionary-based procedure and extracted into location-specific tokens.[19] Finally, we flag person names and professional titles, such as "DR." and "MED.", which may otherwise affect matching quality.[20]

The output of this step is a harmonized set of name tokens for each *crefo*. This preprocessing is central to the name-based pillar because it allows the matching procedure to distinguish informative name components from legal-form, location, and person-name elements that may otherwise generate spurious similarity.

4.4.2 Special Cases Excluded from Fuzzy Matching

Not all business names are suitable for fuzzy matching. Very short or generic names can generate high similarity scores for reasons that are not informative about firm identity. To reduce false positives, we therefore exclude a set of special cases from the fuzzy matching procedure and handle them separately through exact string matching.

First, we restrict the fuzzy matching sample to legal units with common corporate legal forms, including *GmbH*, *OHG*, *KG*, *KGaA*, *UG*, *eG*, *EWG*, and other standard corporate bodies. We exclude legal forms and entity types that are unlikely to form part of larger enterprise structures, including freelancers, partnerships of freelancers, sole proprietors, and associations.

Second, among the remaining names, we exclude strings that are too short or too generic to produce discriminative similarity scores. For example, names consisting of three tokens are

[19] To construct the location dictionary, we select large cities from "Onlineprodukt Gemeinden 31.12.2023" and supplement it with information on the Berlin metropolitan area.

[20] The person-name dictionary contains both first names and surnames and is constructed from common Orbis shareholder names rather than from a complete name list.

excluded if all three tokens contain fewer than three characters, if their combined length is below six characters, or if the first two tokens contain fewer than four characters each.[21] Business names dominated by generic legal suffixes, common abbreviations, or numeric identifiers are also excluded.

Excluded observations are not discarded. Instead, they are matched exactly by standardized string identity within each cross-section. Exact matches among excluded names are then added to the set of name-based links obtained from fuzzy matching. This approach limits Type I error from spurious fuzzy matches among generic names, while retaining reliable exact matches where they exist. Potential Type II error from excluding true but non-exact name links is mitigated by the ownership-based pillar, which can still assign such units to the same firm when common control is observed.

4.4.3 Fuzzy Name Matching with RECLINK2

All harmonized business names not excluded by the rules above enter the fuzzy matching stage. To reduce the number of pairwise comparisons, we first apply phonetic blocking within each year. Specifically, we divide the list of *crefo* names into 100 subsets using a `Soundex`-based blocking key derived from the first token.[22] Within each block, two copies of the data are created to implement exhaustive within-block matching. Missing tokens, up to four tokens per name, are replaced with unique placeholders.[23]

Fuzzy matching is implemented using RECLINK2 in Stata. The algorithm compares up to four word tokens and the phonetic key, assigning higher weights to earlier tokens. We classify two names as a match if their similarity score is at least 0.85.[24] Pairwise matches above this threshold are then grouped into connected components, which define annual name-based clusters.

Figure 6 illustrates the grouping logic. A *crefo* can enter a name-based cluster through either a direct or an indirect link. Direct links arise when two names have a similarity score of at least 0.85. Indirect links arise when two units do not meet the threshold with each other but are each sufficiently similar to a common third unit. As a result, all units connected through the network of above-threshold name matches are assigned to the same annual name-based cluster. Unmatched records form singleton clusters.

[21] Names with only one or two tokens are also excluded under the following conditions: the first token has fewer than three characters and the second token, if present, has more than four; the combined length of both tokens is below six characters; or the first token has fewer than four characters and no second token exists.

[22] Blocking restricts comparisons to names with similar-sounding first tokens, improving computational feasibility and reducing comparisons between implausible candidate pairs.

[23] Placeholders differ across the two copies to prevent false matches on empty fields.

[24] The phonetic key is used as an `orblock` variable to restrict comparisons within blocks.

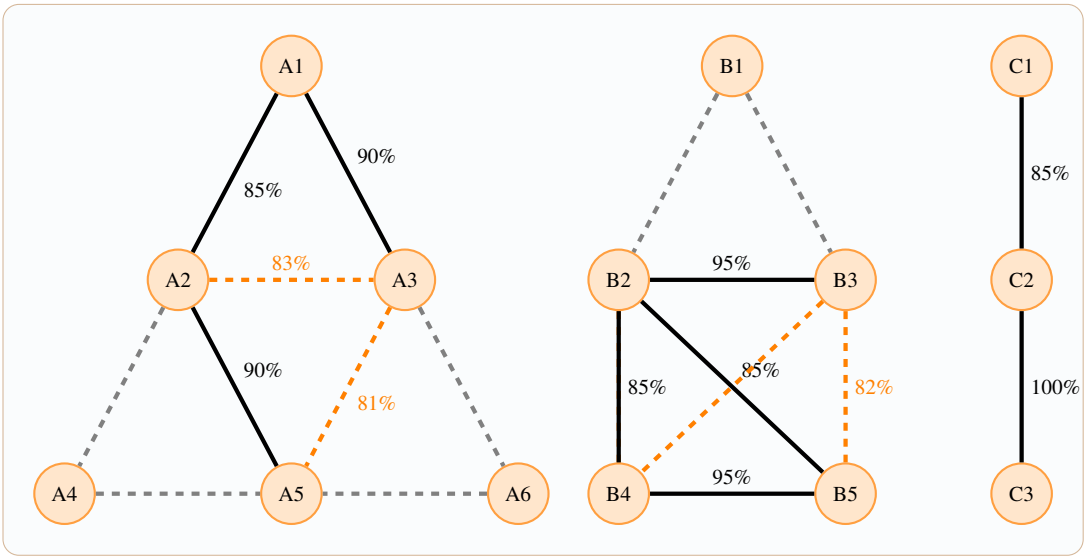


Figure 6: Clustering Based on Business-Name Similarity

Notes: The figure illustrates how pairwise name matches are grouped into annual name-based clusters. Solid lines denote direct fuzzy matches with a similarity score of at least 0.85. Orange dashed lines denote indirect links: two legal units may have a pairwise similarity score below 0.85 but still belong to the same cluster because both are directly matched to a common third unit. Gray dashed lines denote pairwise similarities below the threshold that do not generate a direct match. The algorithm assigns all directly or indirectly connected units to the same name-based cluster. Appendix Table A.2 reports the same example in tabular form.

4.4.4 Longitudinal Identification: Jaccard Similarity

After constructing annual name-based clusters, we impose temporal consistency using the Jaccard similarity routine described in Section 4.2. For name-based clusters, we use a threshold of 0.60. The lower threshold relative to the ownership-based pillar reflects the fact that name-based clusters are more sensitive to changes in recorded names, spelling variation, and the entry or exit of individual *crefos* from the name-matching network.

For each pair of adjacent years, we compare name-based clusters by their *crefo* composition. If the most similar cluster in year $t + 1$ exceeds the threshold, the name-based identifier is retained. If the overlap falls below the threshold, a new identifier is assigned. This procedure allows the name-based identifier to remain stable through minor changes in cluster composition while recognizing substantive changes in the set of legal units connected by business-name similarity.

The output of this step is a time-consistent *Name ID*. This identifier captures continuity in business-name-based clusters and remains available alongside the ownership-based identifier.

4.5 Interpreting the Two Pillars Jointly

The preceding steps produce two time-consistent identifiers: an ownership-based identifier and a name-based identifier. We keep these outputs separate in the data environment because they

capture distinct concepts. The *Ownership ID* is anchored in observed control relations. The *Name ID* is anchored in similarity of business names and can therefore capture links that are not visible in the ownership data. Comparing the two identifiers provides useful information about the nature of the links they identify and about cases in which one source should discipline the other.

The comparison is performed at the *crefo*-year level. Ownership-based links identify legal units under common control, including units that operate under different business names. Name-based links identify legal units with identical or highly similar names, including units for which ownership information is missing, incomplete, or insufficient to establish common control. The rules below are therefore best understood as guidelines for interpreting the two pillars jointly, not as a mechanical construction of a single final identifier.

The comparison between ownership-based and name-based assignments yields three relevant cases.

Case 1: Agreement between ownership and name information. A set of *crefos* is assigned to the same cluster by both procedures. This case provides the strongest evidence that the two data sources identify the same underlying firm structure.

Case 2: Common ownership but different names. A set of *crefos* is assigned to the same ownership-based cluster but to different name-based clusters. This pattern is expected in multi-legal-unit enterprises. Subsidiaries, holding companies, production units, and administrative units may operate under distinct names even though they are subject to common control. In such cases, the ownership pillar identifies a firm structure that name-based clustering alone would miss.

Case 3: Similar names but different ownership clusters. A set of *crefos* is assigned to the same name-based cluster but to different ownership-based clusters. This pattern is informative but ambiguous. It may reflect legally independent firms that operate under a common brand, as in franchise structures. It may also arise when firms share generic or standardized name components. Conversely, similar names may capture economically meaningful continuity, for example when a business name is retained after a sale or restructuring. For this reason, name similarity alone should not be interpreted as common control when ownership information indicates distinct controlling units.

These cases clarify how the two pillars should be interpreted. The ownership pillar is the preferred source for common control when ownership information is available. The name pillar provides additional information where ownership data are missing or incomplete, but it should be used cautiously when similar names coexist with distinct ownership clusters. Keeping both identifiers available allows researchers to choose the level of aggregation appropriate for a given application and to conduct sensitivity analyses around the treatment of ambiguous name-based links.

As a guideline, any combined use of the two pillars should start from the conceptual distinc-

tion between control and name similarity. Legal units connected through common ownership should be interpreted as part of the same control structure even when their names differ. By contrast, name-based links across distinct ownership clusters should not be interpreted as evidence of common control. Name-based links are most informative for units without ownership information or for cases in which ownership data are insufficient to identify control.

Legal units for which neither pillar identifies a link to another unit remain singleton observations in the corresponding identifier. Retaining these singletons is essential: the absence of an observed link does not imply measurement error. It may indicate an independent firm, or a firm for which no broader group structure is observed in the available ownership and name records. This design preserves the full legal-unit population while identifying the subset of units that belong to broader firm structures.

The comparison between the two pillars helps balance Type I and Type II errors. Using ownership information to discipline name-based links reduces Type I error by limiting false merges of legally distinct firms that share similar names. Retaining ownership-based links across different names and allowing name-based links where ownership information is missing reduces Type II error by limiting false splits of legal units that plausibly belong to the same firm. The two-pillar structure therefore preserves the additional coverage provided by the MUP–Business Name Panel without forcing name similarity to imply common control.

The output of the algorithm is therefore a longitudinal *crefo*-year crosswalk containing both the *Ownership ID* and the *Name ID*. This crosswalk can be linked to the MUP–LBD and MUP–BHP to restrict attention to active legal units and to describe the economic relevance of the constructed identifiers. The empirical overlap between the two pillars is examined in Section 5.

5 Disambiguation Results

This section documents the empirical properties of the disambiguation outputs. It describes their coverage, internal structure, and the extent to which the ownership-based and name-based pillars identify overlapping firm structures. The analysis proceeds in three steps. First, we describe the ownership-based output. Second, we examine the coverage and structure of the identifiers among legal units linked to the MUP–LBD and observed in the commercial business register. Third, we describe the overlap between the separately constructed ownership-based and name-based pillars.

Throughout this section, it is important to distinguish algorithmic coverage from economic activity. The raw ownership and business-name inputs accumulate over time and therefore do not represent an annual stock of active firms. Upward trends in raw algorithmic outputs should therefore not be interpreted as evidence of firm entry, consolidation, or changes in organizational complexity. By contrast, analyses based on the MUP–LBD restrict attention to legal units observed as active in a given year. We therefore interpret results based on raw algorithmic inputs

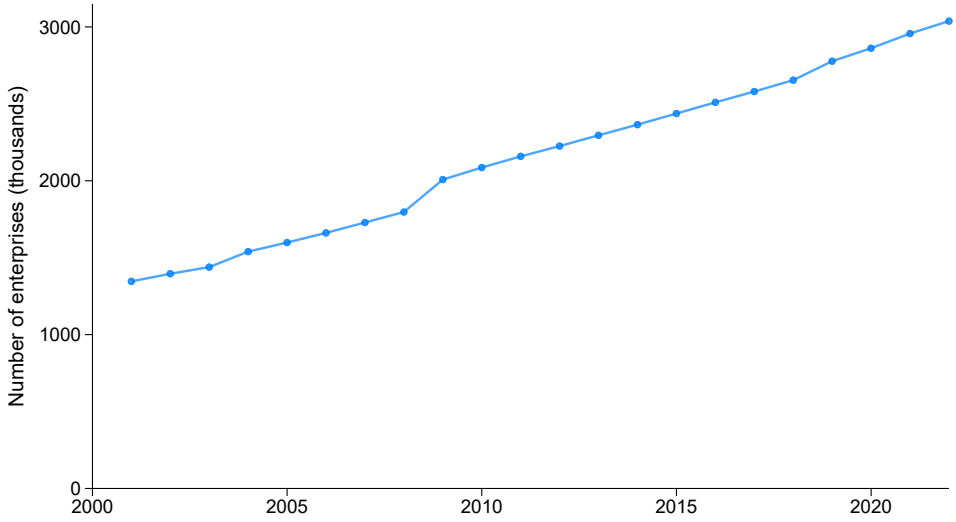


Figure 7: Number of HCU *Crefos*, 2000–2022

Notes: The figure displays the annual number of *crefos* identified as highest controlling units (HCU) based on hierarchical ownership information in the MUP. The sample includes all legal units that have appeared in the ownership data up to the respective year. As a result, the number of HCU *crefos* increases mechanically over time with expanding data coverage. Units without an identified majority owner remain singleton ownership-based clusters within this framework.

as evidence on data coverage and clustering structure, and results based on the MUP–LBD as evidence on the active business population.

Throughout this section, the terms single-unit and multi-unit refer to the size of the relevant algorithmic cluster in a given year. The SLU/MLU terminology is used below for the application that classifies enterprise-years using the ownership-based enterprise identifier.

5.1 Algorithmic Outputs

Figure 7 presents the cross-sectional ownership-based clustering results based on hierarchical control, as described in Section 4.3.1. Specifically, it reports the number of *crefos* in the MUP identified as highest controlling units. The number of HCU *crefos* increases from about 1.35 million in 2001 to about 3 million in 2022, reflecting the accumulation of ownership information in the data. By construction, the sample includes all legal units that have appeared in the ownership data up to the respective year, so the number of identified HCU *crefos* grows mechanically over time.[25]

Figure 8 reports the corresponding cross-sectional output of the name-based pillar, separately

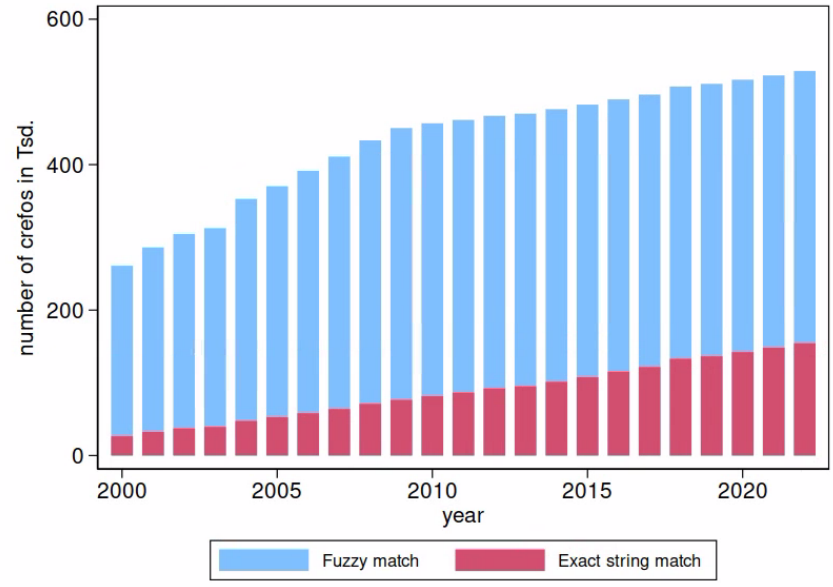
[25] The extent to which changes in the active firm population reflect entry, consolidation, or changes in organizational complexity is assessed in subsequent sections. To assess temporal consistency, Appendix Figure A.1 compares the annual number of newly assigned ownership-based enterprise identifiers under the Jaccard-only, HCU-only, and combined Jaccard and HCU procedures. The combined approach closely follows the HCU rule while correcting selected cases in which changes in the highest controlling owner would otherwise generate a new identifier despite substantial continuity in the underlying ownership structure.

for the full set of *crefos* in Panel A and for the subset observed in the commercial business register in Panel B. It distinguishes legal units linked through fuzzy name matching from those linked through exact standardized-name matches. In both samples, name-based coverage increases over time, with fuzzy matches accounting for the larger component throughout.[26]

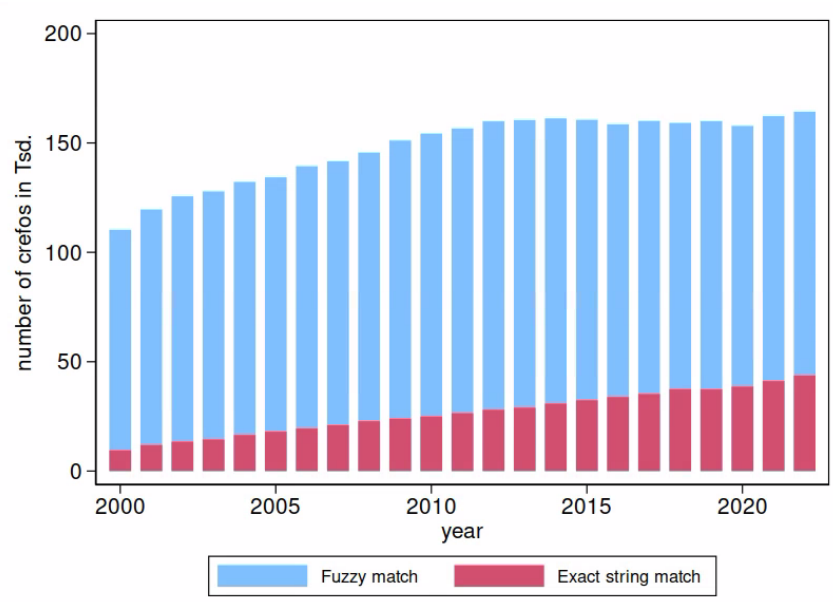
Fuzzy matching identifies the majority of name-based links. In recent years, it accounts for approximately 65–70% of all name-based matches, while exact string matching among excluded names accounts for the remaining 30–35%. The relative contribution of fuzzy and exact matching is similar among commercial-register units, indicating that the same division between fuzzy name matching and exact matching among excluded names also applies to registered legal units.

Figures 7 and 8 document the cross-sectional output of the two algorithmic pillars. The overlap and complementarity between ownership-based and name-based identifiers are examined below.

[26] The name-based matching algorithm is being refined. Preliminary checks suggest that roughly 5% of matching entries exhibit potentially problematic name structures, without necessarily being incorrect matches. The reported number of name matches may therefore be slightly overstated.



(a) All *crefos*



(b) Commercial business register

Figure 8: Cross-Sectional Name-Based Matches

Notes: The figure reports annual cross-sectional name-based matching results using the MUP–Business Name Panel. Panel A uses all available *crefos*; Panel B restricts the sample to *crefos* observed in the commercial business register. Bars distinguish *crefos* linked through fuzzy name matching from *crefos* linked through exact standardized-name matching. Counts are reported in thousands. The figure describes coverage and matching structure in the raw name-based input and should not be interpreted as an annual stock of active legal units or enterprises.

5.2 Coverage among Active Legal Units in the MUP–LBD

We now restrict attention to *crefos* linked to the MUP–LBD. This restriction allows us to describe the clustering results for legal units observed as active in the MUP–LBD.

Figure 9 reports the share of MUP–LBD-linked *crefos* assigned to multi-unit clusters, separately for the ownership-based and name-based pillars.

The ownership-based share increases strongly over time, from about 11% at the beginning of the sample to about 23% by 2022. The name-based share is smaller throughout the period and increases more moderately, from about 3% in 2001 to about 5.5% by 2022. This pattern is consistent with the role of name-based clustering as a complementary component of the algorithm: it identifies additional links based on business-name similarity, but consolidates fewer legal units than the ownership-based pillar in the MUP–LBD-linked sample.

The increase in the ownership-based share reflects two features already documented in Section 3. First, the total number of legal units in the MUP–LBD declines after the late 2000s. Many of the units leaving the active LBD population are small or unregistered units for which ownership information is typically unavailable. Second, the number of registered legal units, and with it the number of units for which ownership information is available, increases over time. As a result, ownership-based clustering applies to a growing share of the active legal-unit population.

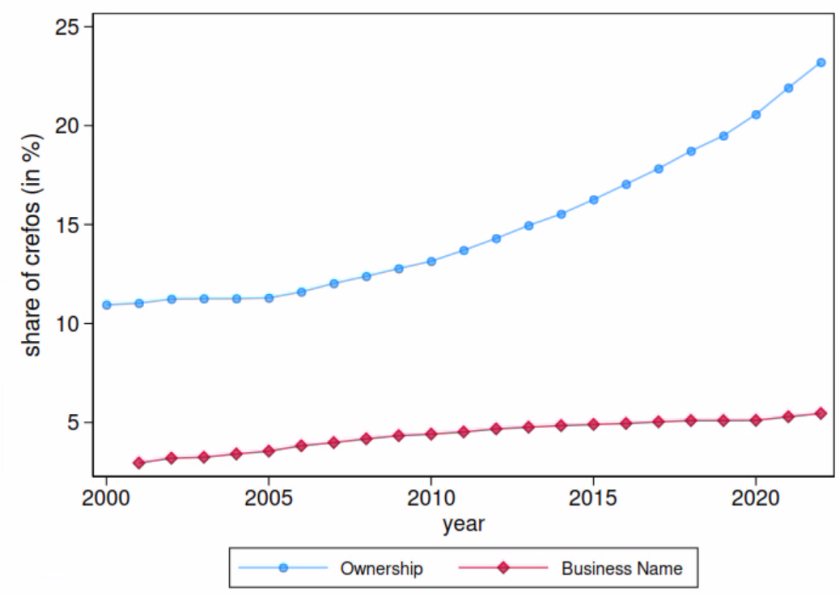


Figure 9: Share of *Crefos* in Multi-Unit Clusters in the MUP–LBD

Notes: The figure reports the share of Creditreform legal units (*crefos*) linked to the MUP–LBD that are assigned to ownership-based or name-based clusters with two or more legal units. The ownership-based and name-based series are separate outputs of the two algorithmic pillars and are not mutually exclusive.

Figure 9 is descriptive. It shows where each pillar changes the unit of observation relative to

individual legal units. The two series should not be interpreted as mutually exclusive categories: a given *crefo* may contribute to both the ownership-based and name-based clustering outputs. The explicit overlap between the two approaches is examined below.

5.2.1 Ownership-Based Cluster Size

The preceding figure showed that both pillars assign a nontrivial share of active MUP–LBD legal units to multi-unit clusters, with a larger role for ownership-based clustering. We next describe the multi-unit structures underlying the ownership-based series. These are the cases in which the ownership-based pillar consolidates several legal units into a common firm structure.

Figure 10 reports the annual shares of enterprise identifiers assigned to ownership-based clusters with exactly two, or with three or more, legal units. It presents these shares separately for the MUP–LBD population and for units observed in the commercial business register. In the MUP–LBD population, the share of ownership-based identifiers with two legal units rises from about 1.8% in 2000 to about 4.2% in 2022, while the share with three or more legal units rises from about 0.8% to about 2.2%. Among registered legal units, the corresponding shares are higher, rising from about 5.9% to about 9.0% for two-unit clusters and from about 2.8% to about 4.7% for clusters with three or more units. In both data sources, clusters with two legal units are more common than larger clusters. Thus, the most frequent consolidation performed by the ownership-based pillar links two legal units, while larger ownership-based structures remain less frequent.

The higher shares in the commercial-register population should not be interpreted as evidence of systematically more complex firm organization. They reflect the more selective registered-unit denominator, whereas the MUP–LBD covers the broader LBD universe. Likewise, changes over time primarily reflect evolving data availability and coverage. The figure's purpose is therefore descriptive: it shows that multi-unit ownership structures are present in both legal-unit data sources and that following ownership links is necessary to recover common firm structures consistently across them.

Taken together, Figure 10 complements Figure 9 by documenting the source of the reduction from *crefos* to ownership-based clusters. The ownership-based pillar matters precisely because it identifies cases in which several legal units should be treated as part of the same firm structure.

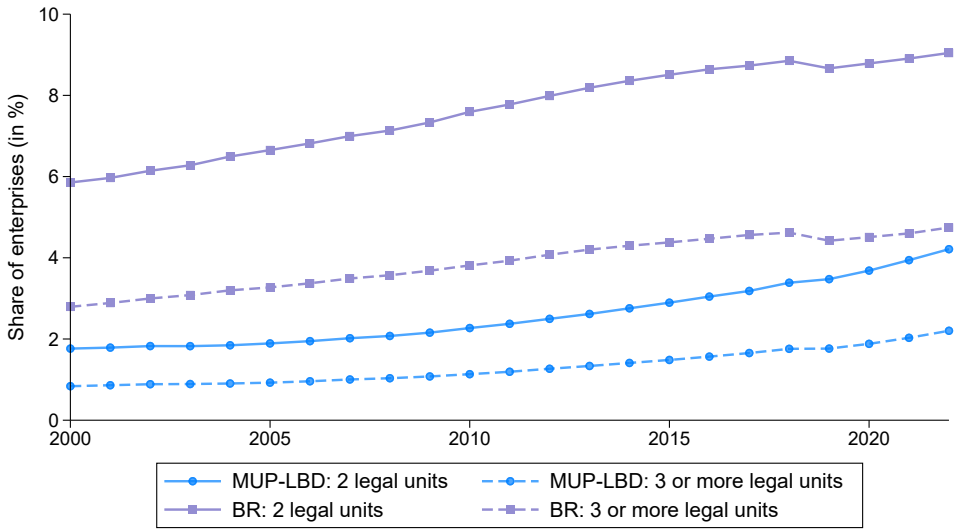


Figure 10: Size Distribution of Multi-Unit Ownership-Based Clusters, 2000–2022

Notes: The figure reports annual shares of ownership-based identifiers assigned to clusters with exactly two legal units or with three or more legal units, separately for the MUP–LBD and commercial-register populations, 2000–2022. Solid lines denote clusters with two legal units; dashed lines denote clusters with three or more legal units. Blue series refer to the MUP–LBD and red series to the commercial business register.

5.3 Overlap between Ownership and Name Clusters

We next compare the ownership-based and name-based clustering outputs among *crefos* observed in the commercial business register. This comparison builds on Figure 9, which documents the two algorithmic pillars separately for LBD-linked units. The figures in this subsection instead describe how the two pillars overlap among registered legal units. They do not impose a final combined identifier.

Figure 11 presents two related decompositions. Panel A reports the annual number of registered *crefos* and classifies them according to the clustering information available from the two pillars. The panel distinguishes four mutually exclusive categories: *crefos* covered only by ownership-based clustering, covered only by name-based clustering, covered by both approaches, and covered by neither approach. Single-unit clusters are included in the first three categories. Hence, the "neither" category reflects cases with neither ownership-based nor name-based cluster information.

Panel B restricts attention to multi-unit cluster assignments. This is the subset in which the algorithm changes the observational unit relative to individual legal units. The panel distinguishes registered *crefos* assigned to multi-unit clusters by both approaches, by ownership-based clustering only, by name-based clustering only, or by neither approach. The category "neither" includes *crefos* that are not covered by any clustering method as well as *crefos* that are covered only by single-unit clusters.

Panel A shows substantial overlap between the two sources among registered legal units. A

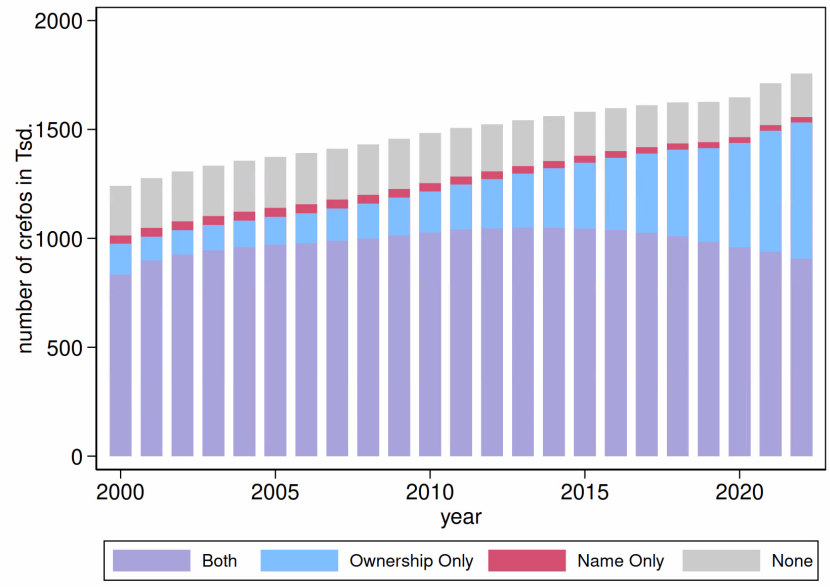
large share of registered *crefos* is covered by both ownership-based and name-based clustering. At the same time, the ownership-only category increases over time, while the name-only category remains comparatively small. Panel B shows that ownership-only assignments account for the largest group of multi-unit links, while name-only assignments provide additional, but smaller, coverage. The overlap category is comparatively small, indicating that the two pillars often identify different multi-unit structures. This pattern is consistent with their distinct information content: ownership-based clustering identifies common control, while name-based clustering identifies similarity in business names.

Figure 12 provides a stricter concordance check based on the agreement between the two pillars. Panel A focuses on singleton assignments. For each pillar, it reports the share of *crefos* classified as singleton observations that are also classified as singleton observations by the other pillar. Panel B focuses on multi-unit assignments. For each pillar, it reports the share of multi-unit cluster assignments with a corresponding cluster in the other pillar, using Jaccard similarity above 50% and conditioning on the availability of both ownership and business-name information.

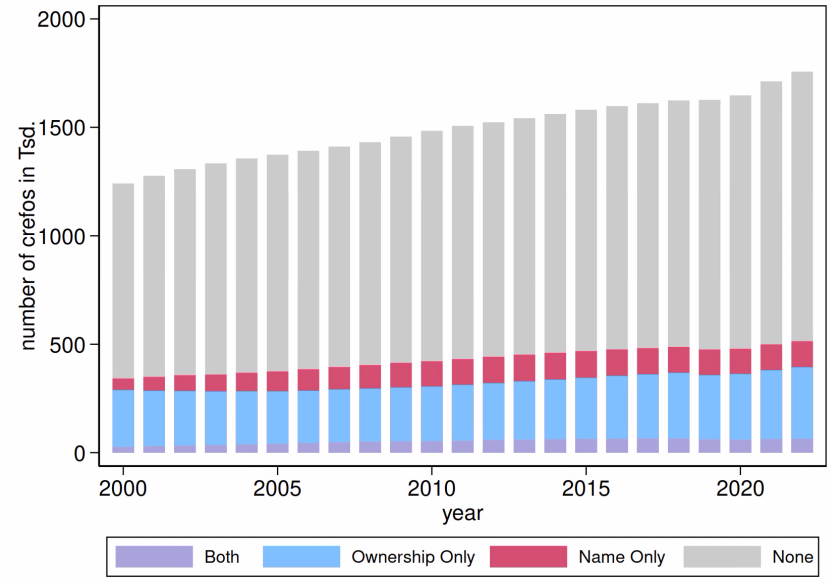
Panel A shows high concordance for single-unit assignments. For ownership-based singletons, the corresponding share remains close to 90% throughout the sample period. For name-based singletons, it declines from about 70% in 2000 to about 62% in 2022, consistent with some legal units that appear as stand-alone entities in the name-based pillar belonging to larger ownership structures. Panel B shows that concordance among multi-unit assignments is much lower, but increases over time. The corresponding share rises from below 1% to about 2.3% for ownership-based multi-unit enterprises, and from about 1.1% to about 7.3% for name-based multi-unit enterprises.

The asymmetry is consistent with the conceptual distinction between the two pillars. Legal units with similar names are more likely to map into a common ownership structure than ownership-linked legal units are to share similar names. Ownership groups often contain subsidiaries, holding entities, or functionally distinct units with different names. Conversely, similar names may also occur across separately controlled entities, for example in franchises, common brands, or generic business-name conventions.

Taken together, these descriptives establish where the two pillars provide distinct and overlapping information. The next section uses the linked MUP–BHP to examine how the resulting identifiers change measured firm structure and economic aggregates.



(a) All cluster assignments



(b) Multi-crefo cluster assignments

Figure 11: Overlap of Ownership- and Name-Based Clusters among Registered *Crefos*

Notes: Panel A reports the annual number of *crefos* observed in the commercial business register and their coverage by the ownership-based and name-based clustering procedures. Categories are mutually exclusive and distinguish *crefos* covered only by ownership-based clustering, only by name-based clustering, by both approaches, or by neither approach. Single-unit clusters are included in the first three categories, such that the "neither" category reflects missing ownership-based and missing name-based cluster information. Panel B reports the corresponding decomposition for multi-unit cluster assignments. In Panel B, the "neither" category includes *crefos* not covered by any clustering method and *crefos* covered only by single-unit clusters. Counts are reported in thousands.

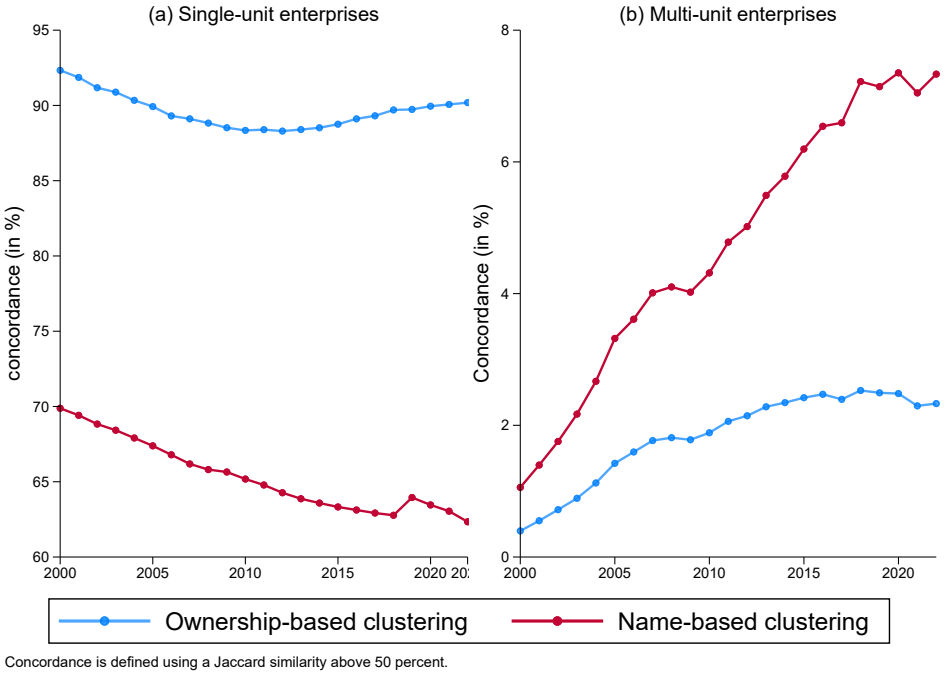


Figure 12: Concordance between Ownership-Based and Name-Based Clustering

Notes: Panel A reports, separately by pillar, the share of *crefos* in the commercial business register classified as singleton observations that are also classified as singleton observations by the other pillar. Panel B reports, separately by pillar, the share of multi-unit cluster assignments with a corresponding cluster in the other pillar, using Jaccard similarity above 50%. Both panels condition on available ownership and business-name information. Shares are reported in percent.

6 Economic Relevance: Applications of the Enterprise Identifier

6.1 Employment Coverage and Enterprise Structure in the MUP–BHP

We finally link the ownership-based enterprise identifiers to the MUP–BHP to assess their economic relevance in employment terms. The linked data contain three distinct observational levels. The BHP records employment by administrative establishment identifier (*betnr*); the MUP–IAB crosswalk links these BHP establishments to *crefo* legal units; and our ownership-based identifier groups *crefo* legal units into enterprises. The first mapping need not be one-to-one, and a BHP establishment may coincide with its legal unit in simple organizations. Throughout the application, however, we reserve *establishment* for the BHP employment-reporting unit, *legal unit* for the *crefo*, and *enterprise* for the ownership-based cluster.

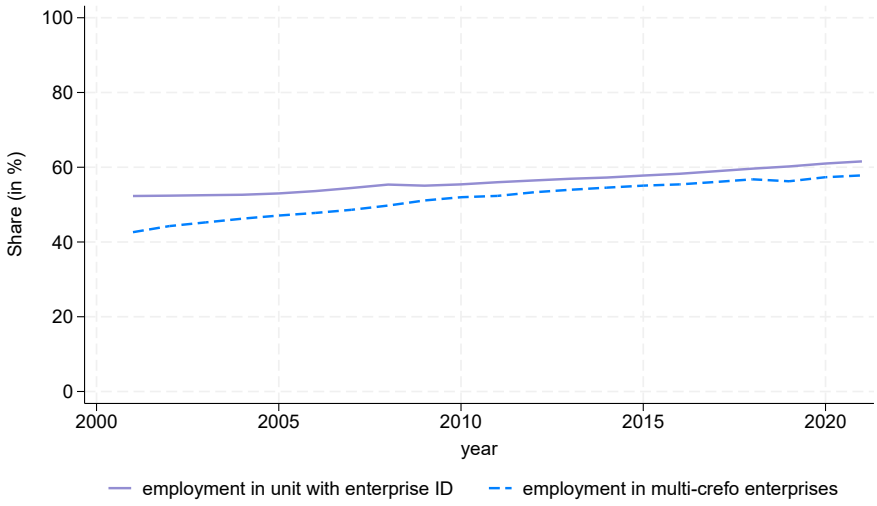


Figure 13: Employment Coverage of Enterprise Identifiers in the MUP–BHP

Notes: The figure reports annual employment shares. The solid line shows the share of overall employment in Germany, measured in the BHP, that can be assigned to a time-consistent ownership-based enterprise identifier through the MUP–BHP linkage. The dashed line shows, among employment covered by an enterprise identifier, the share located in MLU enterprises.

The SLU/MLU distinction refers exclusively to the number of legal units within an enterprise and not to its number of establishments. An SLU enterprise may be a single-establishment enterprise or may operate several BHP establishments through its single legal unit. Conversely, an MLU enterprise contains several legal units even if each legal unit is linked to only one BHP establishment. This distinction matters because MLU enterprises may be a minority of units while accounting for a large share of workers.

Figure 13 reports two employment shares. First, around 55–60 percent of overall employment in Germany, as measured by the BHP, can be assigned to an ownership-based enterprise identifier through our algorithm. Second, among employment covered by an enterprise identifier, roughly 42–58 percent is located in MLU enterprises. In levels, this corresponds to employment in MLU enterprises rising from about 6 million workers at the beginning of the sample to more than 12 million workers by the end, or slightly more than 30 percent of total BHP employment in the final years. Thus, the enterprise identifier covers a large part of aggregate employment, and a substantial share of that coverage lies in enterprises that contain multiple legal units.

Illustrative organizational forms that may generate an MLU structure include a corporate group with separately incorporated operating subsidiaries, a holding company controlling several production or service companies, or an organization that places regional, production, distribution, and administrative activities in distinct legal entities. These examples are conceptual rather than descriptions of particular firms in the confidential data. They clarify why a legal-unit identifier can divide an economically coordinated organization into several observations.

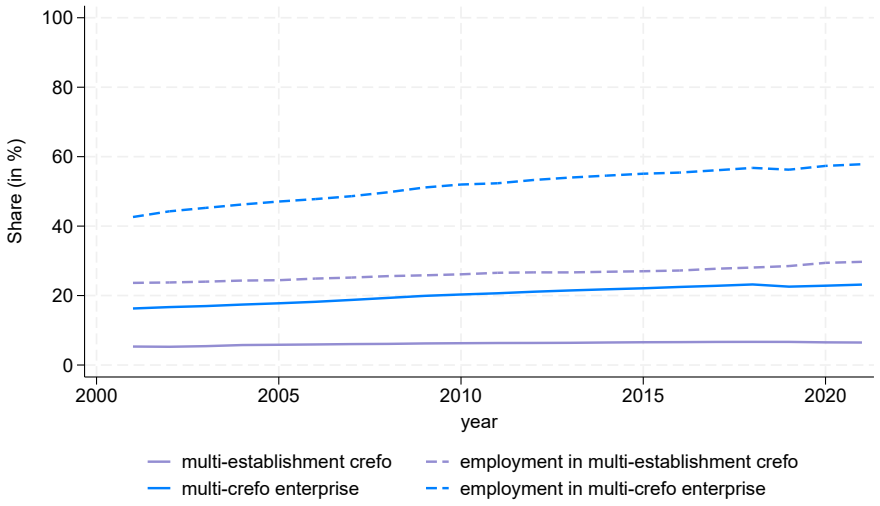


Figure 14: Legal-Unit and Enterprise Structure in the MUP–BHP

Notes: The figure describes organizational structure in the MUP–BHP. Multi-establishment *crefos* are legal units linked to more than one establishment in a given year. MLU enterprises are enterprise-years containing more than one *crefo* legal unit. The figure reports both unit shares and employment-weighted shares. *Crefo* shares are computed relative to all linked *crefos*, except MLU-enterprise shares, which are computed among *crefos* with an ownership-based enterprise-identifier match. Employment shares use the corresponding employment totals as denominators.

Figure 14 decomposes the relevant aggregation margins. Only around 5–7% of *crefo* legal units are multi-establishment legal units, but they account for roughly 23–30% of employment covered by the MUP–BHP linkage. By contrast, around 16–23% of *crefo* legal units with enterprise-identifier coverage belong to MLU enterprises. These enterprises account for a much larger employment share: roughly 42–58% of employment covered by an enterprise identifier, increasing over time. Combining this pattern with the overall employment coverage of enterprise identifiers implies that MLU enterprises account for about 20% of total BHP employment at the beginning of the sample and more than 30% by the end.

The figures show that the enterprise identifier changes the empirical unit in an economically meaningful way. The issue is not primarily that some *crefo* legal units contain several establishments. That margin is visible but relatively small in unit counts. The larger issue is that many *crefo* legal units belong to MLU enterprises, and these enterprises account for a disproportionately large and increasing share of employment. Hence, legal-unit analysis can split economically integrated firms into separate observations precisely where employment is most concentrated.

The enterprise identifier is therefore most consequential in settings where considering the firm as the decision making unit of control is relevant, and in particular whenever firm size, concentration of activity, ownership change, restructuring, reallocation or innovation are central objects of analysis. It provides an ownership-based approximation to the relevant enterprise struc-

ture and moves the unit of analysis closer to the organization in which employment, investment, and ownership decisions are made.

6.2 Enterprise Structure and Measured Job Reallocation

Job creation and job destruction are canonical measures of firm dynamism. In the BHP, job creation arises when an establishment identifier first appears with positive employment or when a continuing establishment expands. Job destruction arises when a continuing establishment contracts or when an establishment identifier disappears from the panel. Gross job reallocation is the sum of these two margins. These are establishment-level accounting objects. Their interpretation as firm dynamics, however, depends on the legal-unit and enterprise structures in which the establishments are embedded.

As Haltiwanger, Jarmin, and Miranda (2013) emphasize, measuring firm dynamics requires following both establishments and the firms to which they belong over time. Employment growth at a continuing establishment records additional jobs at that unit. Changes in a firm's employment total, however, can also reflect mergers, acquisitions, or divestitures that transfer existing establishments across firm boundaries. When employment growth is attributed to firm characteristics, such as size or age, these boundary changes must be distinguished from employment changes at the underlying establishments. Our application follows this measurement logic by calculating job flows at the BHP establishment level and then classifying them by enterprise structure. For a continuing BHP establishment with unchanged employment, a change in enterprise assignment contributes neither job creation nor job destruction. The linkage nevertheless records the organizational structure to which the establishment belongs. Joint observation therefore combines establishment-level employment accounting with information on enterprise boundaries, which neither observational level provides on its own.

Enterprise information is also needed because establishment-level job flows can arise in different organizational contexts. A newly observed BHP establishment may represent a new stand-alone enterprise, an additional establishment of a continuing legal unit, or an establishment attached to a newly created legal unit within an existing enterprise. The same ambiguity applies to job destruction: the disappearance or contraction of a BHP establishment may reflect enterprise decline or exit, but it may also occur within a continuing legal unit or a continuing multi-legal-unit enterprise. BHP establishment data alone do not distinguish these contexts. Linking each establishment through its *crefo* legal unit to an enterprise boundary provides that organizational information, although it does not by itself identify the mechanism generating a particular employment change.

The enterprise identifier makes this organizational classification observable. We first calculate employment changes separately for each BHP establishment. We then classify each establishment-year using the enterprise structure of its linked *crefo*. An establishment is assigned to the identified MLU category when it is linked to a legal unit in an enterprise containing more

than one legal unit. All remaining establishment-years, including those without an identified *crefo* or enterprise match, form the residual SLU category. The two categories therefore provide an exhaustive decomposition of the full BHP job-flow totals, while only the MLU component requires an affirmative enterprise-linkage match.

Importantly, this decomposition does not change the aggregate job creation, job destruction, or gross reallocation rates. It partitions the same establishment-level employment changes according to whether the establishment is embedded in an identified MLU enterprise or belongs to the residual SLU category. Employment is never aggregated to the *crefo* level before job flows are calculated. Consequently, an employment gain at one establishment and an offsetting loss at another establishment remain separate creation and destruction flows even when both establishments belong to the same legal unit or enterprise.

The enterprise identifier adds information about the organizational setting in which these establishment-level flows occur. For establishments linked to MLU enterprises, the observed flow may reflect enterprise-wide expansion or contraction or a reallocation of activity within the broader organization. For establishments linked to identified SLU enterprises, reallocation across other legal units within the same observed enterprise is ruled out. Because unmatched establishments are included in the residual SLU category, that category should be interpreted as a residual rather than as proof that every establishment operates independently. The measurement contribution is to quantify the portion of aggregate job flows for which a broader identified organizational context may matter, without attributing those flows to a particular reorganization mechanism.[27]

[27] The MLU component could be decomposed further into establishment entry and exit, expansion and contraction among continuing establishments, the creation or closure of legal units within continuing enterprises, changes in enterprise composition, or M&A-related boundary changes. The present application leaves these margins aggregated because its purpose is to demonstrate the additional measurement made possible by observing enterprise boundaries.

Formally, let b index BHP establishments, let $\ell_t(b)$ denote the *crefo* legal unit to which establishment b is linked in year t, and let $e_t(b)$ denote the enterprise containing that legal unit. We set the corresponding index to the null label $\varnothing$ when a legal-unit or enterprise link is unavailable. Write $E_{b,\ell_t(b),e_t(b),t} \equiv E_{bt}$ for employment at establishment b. The legal-unit and enterprise indices record the establishment's organizational assignment; they are not levels at which employment is aggregated. The employment change is always formed at the establishment level,

$$\Delta E_{bt} = E_{b,\ell_t(b),e_t(b),t} - E_{b,\ell_{t-1}(b),e_{t-1}(b),t-1} \equiv E_{bt} - E_{b,t-1},$$
$$X_{bt} = \tfrac{1}{2}\left(E_{bt} + E_{b,t-1}\right).$$

For an establishment first observed in year t, $E_{b,t-1}$ is set to zero; for an establishment last observed in $t-1$, E_{bt} is set to zero. Entry and exit therefore refer to the first and last observation of the BHP establishment identifier, not to the incorporation or dissolution of a legal unit or enterprise.

Let $C_{bt} \in \{\text{SLU}, \text{MLU}\}$ denote the enterprise-structure category assigned to establishment b in year t, where MLU requires an affirmative link to an enterprise containing multiple legal units and SLU is the residual category.

Following Davis and Haltiwanger (1992), the group-specific annual rates are

$$JC_t^g = 100 \times \frac{\sum_b \mathbf{1}\{C_{bt} = g\} \max(\Delta E_{bt}, 0)}{\sum_b X_{bt}},$$
$$JD_t^g = 100 \times \frac{\sum_b \mathbf{1}\{C_{b,t-1} = g\} \max(-\Delta E_{bt}, 0)}{\sum_b X_{bt}}, \quad g \in \{\text{SLU}, \text{MLU}\}.$$

All components use the same full-BHP denominator, $\sum_b X_{bt}$, where the sum covers the union of establishments observed in $t-1$ or t. The SLU and MLU components therefore sum to the conventional aggregate job creation and job destruction rates in each year. Positive flows are classified by enterprise structure in year t, whereas negative flows are classified by enterprise structure in year $t-1$. This timing convention assigns employment gains to the organizational structure receiving the employment and employment losses to the organizational structure from which employment disappears. The figures report 2002–2021 because construction of the flows requires employment in $t-1$.

Figure 15 shows how much aggregate establishment-level employment expansion is associated with the residual SLU category and with identified MLU enterprises. The total bar is the conventional BHP job-creation rate. The decomposition reveals whether measured job creation occurs primarily at establishments outside identified MLU structures or whether a substantial part occurs at establishments whose linked legal units are embedded in enterprises operating through multiple legal units. This organizational distinction is not visible in BHP establishment data alone.

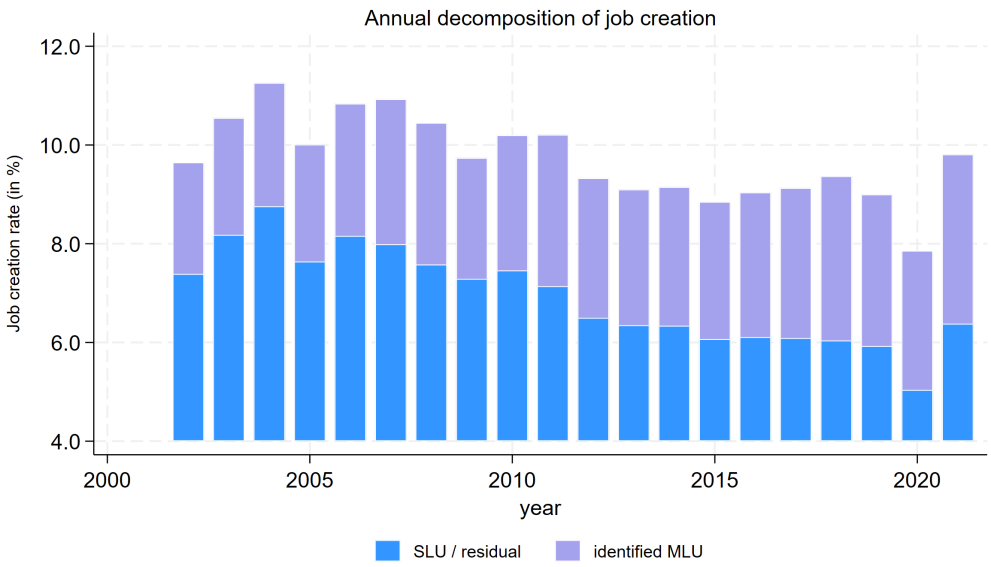


Figure 15: Annual Decomposition of Job Creation by Enterprise Structure

Notes: The figure reports the annual job creation rate calculated from BHP establishment-level employment changes and decomposed by enterprise structure. For each establishment identifier b, job creation is $\max(E_{bt} - E_{b,t-1}, 0)$; an establishment first observed in year t contributes employment growth from zero. Employment is not aggregated to the *crefo* legal-unit level before changes are calculated. Each positive establishment flow is classified using the enterprise structure in year t: identified MLU if the linked legal unit belongs to an enterprise containing multiple legal units, and residual SLU otherwise, including establishments without an identified enterprise match. Components use the same full-BHP average-employment denominator and sum to the aggregate job creation rate in each year.

Figure 16 gives the corresponding decomposition for establishment-level employment contraction. Again, the aggregate job-destruction rate is unchanged. The enterprise identifier reveals whether destruction occurs primarily at establishments outside identified MLU structures or at establishments linked to legal units within MLU enterprises. This matters because the closure or contraction of an establishment embedded in a multi-legal-unit enterprise need not have the same economic interpretation as an otherwise similar flow outside an identified MLU structure.

Figure 17 summarizes the same information for gross job reallocation. It shows the share of establishment-level job-flow activity assigned to the residual SLU category and the share assigned to identified MLU enterprises. The central message is a measurement one: the enterprise identifier does not alter aggregate dynamism, but it reveals an organizational dimension of measured establishment flows that is otherwise hidden. Once enterprise boundaries are observed, reallocation at BHP establishments can be separated according to whether their linked legal units are embedded in identified multi-legal-unit organizations.

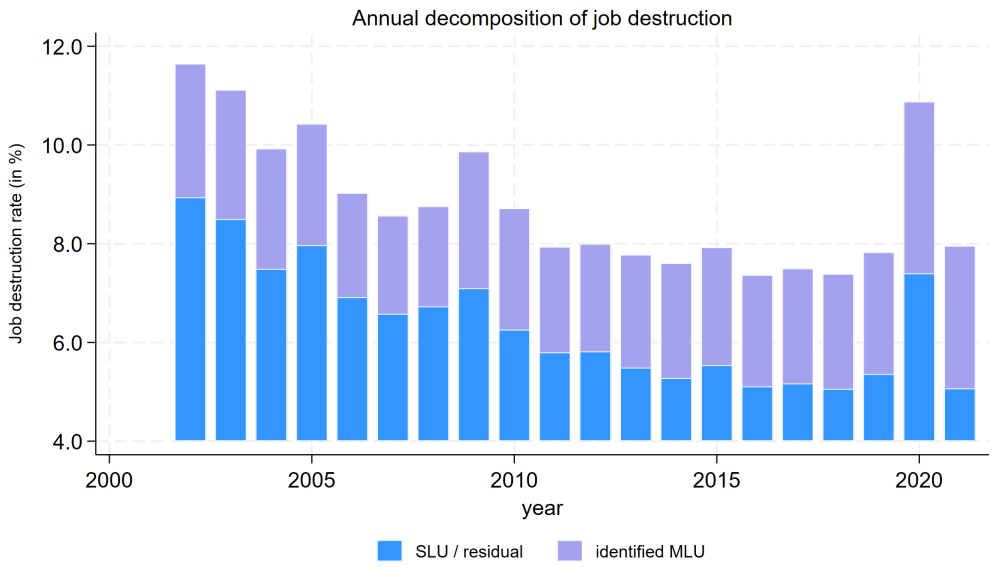


Figure 16: Annual Decomposition of Job Destruction by Enterprise Structure

Notes: The figure reports the annual job destruction rate calculated from BHP establishment-level employment changes and decomposed by enterprise structure. For each establishment identifier b, job destruction is $\max(E_{b,t-1} - E_{bt}, 0)$; an establishment last observed in year $t-1$ contributes an employment decline to zero. Employment is not aggregated to the *crefo* legal-unit level before changes are calculated. Each negative establishment flow is classified using the enterprise structure in year $t-1$: identified MLU if the linked legal unit belongs to an enterprise containing multiple legal units, and residual SLU otherwise, including establishments without an identified enterprise match. Components use the same full-BHP average-employment denominator and sum to the aggregate job destruction rate in each year.

This decomposition also clarifies what richer applications of the enterprise identifier can do. The same hierarchy can distinguish establishment entry within a continuing legal unit from the creation of a new legal unit, and it can distinguish either event from genuine enterprise entry. Analogous distinctions apply to establishment, legal-unit, and enterprise exit. Enterprise identifiers can also be combined with ownership-change data to isolate M&A-related boundary changes. The present figures leave these margins aggregated. Their purpose is to show, using a canonical establishment-level job-flow statistic, what becomes visible once BHP establishments are linked through legal units to enterprise boundaries.

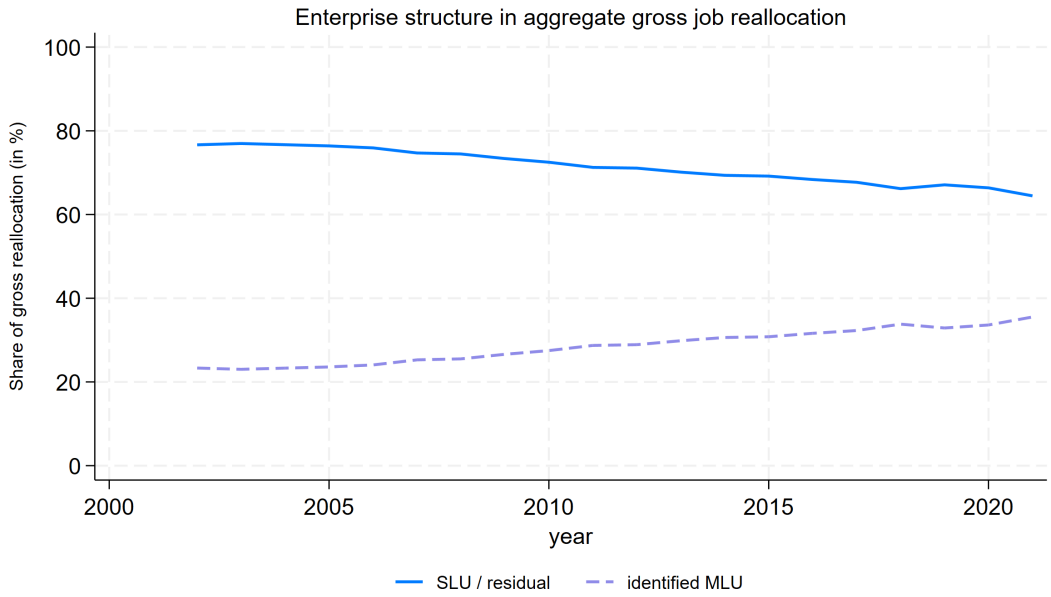


Figure 17: Enterprise Structure in Aggregate Gross Job Reallocation

Notes: The figure reports the share of aggregate gross job reallocation accounted for by BHP establishment flows in the residual SLU and identified MLU categories. Gross reallocation is the sum of establishment-level job creation and job destruction in year t. For group g, the plotted share is $(JC_t^g + JD_t^g)/\sum_h (JC_t^h + JD_t^h)$, computed from the underlying employment-flow components. Job creation is classified by enterprise structure in year t, and job destruction by enterprise structure in year $t-1$. The figure does not separately decompose establishment entry and exit, legal-unit creation and closure, internal reorganization, or M&A events. The two components sum to one in each year.

7 Concluding Remarks

This paper develops a methodology for constructing time-consistent firm identifiers from legal-unit records in the Mannheim Enterprise Panel (MUP) and applies the resulting ownership-based enterprise identifier in the linked MUP–BHP. The underlying measurement problem is fundamental for empirical research using business data. Administrative and commercial sources typically observe establishments or legal units, while many economic questions concern firms as coherent decision-making entities. Without consistent identifiers at this level, analyses of firm dynamics, ownership structures, productivity, innovation, labor-market outcomes, and business networks risk being conducted at an inappropriate level of aggregation.

To address this challenge, we develop two complementary pillars. The ownership-based pillar uses majority-control relations to identify legal units under common control and applies hierarchical control rules together with Jaccard similarity to construct longitudinally consistent ownership structures. The name-based pillar uses harmonized business names, fuzzy matching, and exact matching among names unsuitable for fuzzy matching to identify additional links between legal units. A central insight of the paper is that the two sources capture distinct

dimensions of firm structure. Ownership information directly identifies common control but may be incomplete. Name-based information expands coverage where ownership data are missing or incomplete, but it cannot by itself establish common control. Keeping both pillars available makes the identifier environment transparent and allows researchers to distinguish common control from name similarity.

The main outputs are *crefo*-year crosswalks that assign legal units to time-consistent ownership-based and name-based identifiers. These crosswalks can be linked to the MUP–BHP and allow researchers to aggregate legal-unit and establishment-level information to a more economically meaningful firm level while preserving the possibility of analysis at the legal-unit level. The descriptive evidence shows that the resulting identifiers materially alter the observed unit structure relative to raw *crefo* identifiers. While most identified enterprises are SLU enterprises, MLU enterprises account for a substantial share of economic activity and would remain fragmented in analyses conducted solely at the legal-unit level.

The first application to the MUP–BHP demonstrates the coverage and economic relevance of the resulting identifiers. Around 55–60% of total employment in Germany, as measured in the BHP, can be assigned to an ownership-based enterprise identifier through the linkage. Among employment covered by an enterprise identifier, roughly 42–58% is located in MLU enterprises, with this share increasing over time. Combining these two margins implies that MLU enterprises account for about 20% of total BHP employment at the beginning of the sample period and more than 30% by the end.

These patterns clarify why firm disambiguation matters for empirical economics. The main measurement challenge is not primarily that individual *crefo* legal units contain multiple establishments: only around 5–7% of *crefos* are multi-establishment legal units, although they account for roughly 23–30% of employment covered by the MUP–BHP linkage. Rather, the larger issue is that many *crefos* belong to MLU enterprises. Around 16–23% of *crefos* with enterprise-identifier coverage belong to such enterprises, and these enterprises account for a disproportionately large share of employment. Legal-unit identifiers therefore split economically integrated firms into separate observations precisely in parts of the economy where employment is highly concentrated. Moving from legal units to time-consistent enterprise identifiers brings the unit of analysis closer to the organization in which employment, investment, ownership, and restructuring decisions are plausibly made.

The job-reallocation application provides one illustration of the same measurement logic. It is not intended to make business dynamism the substantive focus of the paper; rather, it uses a standard outcome from the firm-dynamics literature to show how the enterprise identifier changes the interpretation of familiar statistics. We calculate job creation and job destruction from employment changes at the BHP establishment level and then classify those flows through the establishment–legal unit–enterprise linkage. This accounting does not treat enterprise structure as a causal mechanism. It shows what becomes observable once establishment-level outcomes

can be placed within a time-consistent enterprise boundary. A sizable share of aggregate job creation and destruction is associated with establishments linked to MLU enterprises and would be obscured in data organized only around BHP establishment or legal-unit identifiers.

This decomposition changes the economic content of standard business-dynamism statistics without changing their aggregate values. Job flows are formed separately for each BHP establishment, so offsetting changes across establishments within the same legal unit or enterprise do not cancel before job creation and destruction are measured. The enterprise identifier then reveals which of these flows occur inside organizations operating through multiple legal units. The same establishment-level employment change may have a different interpretation depending on whether it occurs within an identified MLU enterprise or outside such a structure. The identifier therefore improves the organizational interpretation of measured job creation, job destruction, and gross reallocation while preserving the conventional BHP aggregates.

Diegmann, Müller, and Schoefer (2026) demonstrate the importance of distinguishing establishments from legal units in research on wage setting. Our enterprise identifier makes it possible to extend such analyses to the level of common control, including specifications with enterprise-level fixed effects across establishments belonging to different legal units.

More broadly, the resulting data infrastructure facilitates empirical research on German firms by linking legal-unit records, ownership structures, business-name information, and administrative labor-market data within a consistent longitudinal framework. The identifiers enable analyses of firm dynamics, restructuring, mergers and acquisitions, productivity, innovation, business groups, and employment outcomes at a level that more closely approximates the underlying economic decision-making unit. In this sense, the paper contributes both a practical implementation of the enterprise concept in a large-scale longitudinal business database and a transparent methodology that can be adapted to other administrative and commercial data environments.

References

Arndt, Lukas (2023). "When Should We Believe Research Using ORBIS Firm Data?" *Available at SSRN 4584106*.

Arora, Ashish, Sharon Belenzon, and Lia Sheer (2021). "Knowledge Spillovers and Corporate Investment in Scientific Research". *American Economic Review* 111.3, pp. 871–98.

Ascione, Grazia Sveva, Valerio Sterzi, and Andrea Vezzulli (2025). "Terrorizer: a novel algorithm for patent assignee name disambiguation: G. Ascione et al." *Scientometrics* 130.12, pp. 7303–7342.

Bajgar, Matej, Giuseppe Berlingieri, Sara Calligaris, Chiara Criscuolo, and Jonathan Timmis (2020). "Coverage and representativeness of Orbis data". *OECD*.

— (2023). "Industry concentration in Europe and North America". *Industrial and Corporate Change*, dtac059.

Beck, Martin, Luisa Baumgärtner, Katja-Verena Bürk, and Matthias Redecker (2020). "New EU definition of enterprise as statistical unit: theory and implementation". *WISTA Scientific Journal*.

Bremer, Leon (2023). *Fuzzy firm name matching: Merging Amadeus firm data to PATSTAT*. Tech. rep. Tinbergen Institute Discussion Paper.

Calligaris, Sara, Miguel Chaves, Chiara Criscuolo, Joshua De Lyon, Andrea Greppi, and Oliviero Pallanch (2024). "Industry concentration in Europe: trends and methodological insights". *OECD Science, Technology and Industry Working Paper* 6.

Cooper, Michael, John McClelland, James Pearce, Richard Prisinzano, Joseph Sullivan, Danny Yagan, Owen Zidar, and Eric Zwick (2016). "Business in the United States: who owns it, and how much tax do they pay?" *Tax Policy and the Economy* 30.1, pp. 91–128.

Davis, Steven J and John Haltiwanger (1992). "Gross job creation, gross job destruction, and employment reallocation". *The Quarterly journal of economics* 107.3, pp. 819–863.

De Monte, Enrico, Thorsten Doherr, Jakob Ehlich, Lena Füner, Sandra Gottschalk, Simona Murmann, and Hanna Hottenrott (2026, forthcoming). "MUP-LBD: The ZEW Mannheim Enterprise Panel for the Analysis of Longitudinal Business Dynamism". *Journal of Economics and Statistics*.

Diegmann, André, Sandra Gottschalk, Mirja Hälbig, Alexandra Schmucker, Stefanie Wolter, and Florian Zimmermann (2026). "MUP-BHP – Enriching Administrative Establishment Data with Commercial Enterprise Data". *Journal of Economics and Statistics*.

Diegmann, André, Steffen Müller, and Benjamin Schoefer (2026). *Off the Labor Supply Curve: The Zero Employer Size Wage Effect Within Large Firms*. NBER Working Paper 35265. Cambridge, MA: National Bureau of Economic Research.

Doherr, Thorsten (2023). "The SearchEngine: A holistic approach to matching". *ZEW-Centre for European Economic Research Discussion Paper* 23-001.

Eurostat (2024). *European Business Statistics Methodological Manual for Structural Business Statistics: 2024 Edition*. Publications Office of the European Union. Luxembourg.

Garcia-Bernardo, Javier, Jan Fichtner, Frank W Takes, and Eelke M Heemskerk (2017). "Uncovering offshore financial centers: Conduits and sinks in the global corporate ownership network". *Scientific reports* 7.1, p. 6246.

Garcia-Bernardo, Javier and Frank W Takes (2018). "The effects of data quality on the analysis of corporate board interlock networks". *Information systems* 78, pp. 164–172.

Haltiwanger, John, Ron S. Jarmin, and Javier Miranda (2013). "Who creates jobs? Small versus large versus young". *Review of Economics and Statistics* 95.2, pp. 347–361.

Jarmin, Ron S. and Javier Miranda (2002). *The Longitudinal Business Database*. CES Working Paper 02-17. Washington, DC: Center for Economic Studies, U.S. Census Bureau.

Reich, Valentin Ferdinand Michael (2024). *Machine Learning based linkage of company data for economic research: Application to the EBDC Business Panels*. Tech. rep. ifo Working Paper.

Sturm, Roland (2018). "The unit problem from a statistical business register perspective". *The Unit Problem and other Current Topics in Business Survey Methodology; Lorenc, B., Smith, PA, Bavdaž, M., Haraldsen, G., Nedyalkova, D., Zhang, L.-C., Zimmermann, T., Eds*, pp. 19–30.

Wasi, Nada and Aaron Flaaen (2015). "Record linkage using Stata: Preprocessing, linking, and reviewing utilities". *The Stata Journal* 15.3, pp. 672–697.

A Appendix

This appendix collects supplementary material supporting the construction of the firm identifiers. It first documents the longitudinal refinement of ownership-based clusters by combining the HCU approach and the Jaccard similarity approach. It then provides tabular versions of the ownership- and name-based clustering examples discussed in the main text.

A.1 Longitudinal Refinement of Ownership-Based Clusters

A.1.1 Combining Jaccard Similarity and HCU

Figure A.1 compares the annual number of newly assigned ownership-based enterprise identifiers under three alternative longitudinal linking rules: a pure Jaccard approach, a pure HCU approach, and the combined Jaccard+HCU procedure described in Section 4.3.2. For comparability, the figure is restricted to ownership clusters containing at least one legal unit observed in the commercial business register.

The combined Jaccard+HCU procedure should be interpreted as a longitudinal refinement of the ownership-based clustering pillar. Unlike the overlap analysis in the main text, which evaluates consistency between ownership and business-name information, Figure A.1 focuses exclusively on the stability of ownership-based identifiers over time under alternative longitudinal linking rules.

The three approaches differ substantially in the frequency with which they generate new identifiers. The Jaccard-only approach assigns between approximately 250,000 and 350,000 new enterprise identifiers per year, whereas the HCU-only and combined approaches generate roughly 25,000 new identifiers annually. This difference follows directly from the construction of the methods. The Jaccard approach evaluates whether the composition of legal units within a cluster remains sufficiently similar between adjacent years. While this criterion is intuitive, it is particularly sensitive in smaller clusters, where the entry or exit of a single legal unit can substantially reduce the similarity measure and trigger a new identifier. As a result, the Jaccard-only approach generates a large number of cluster breaks that are unlikely to correspond to economically meaningful changes in firm identity.

In contrast, the HCU approach assigns a new identifier only when the highest controlling unit changes. Consequently, it produces substantially more stable identifiers over time. The combined procedure largely follows the HCU approach, as reflected by the close correspondence of the two series throughout most of the sample period. This pattern indicates that ownership hierarchy provides the dominant signal for longitudinal continuity in the ownership-based identifier.

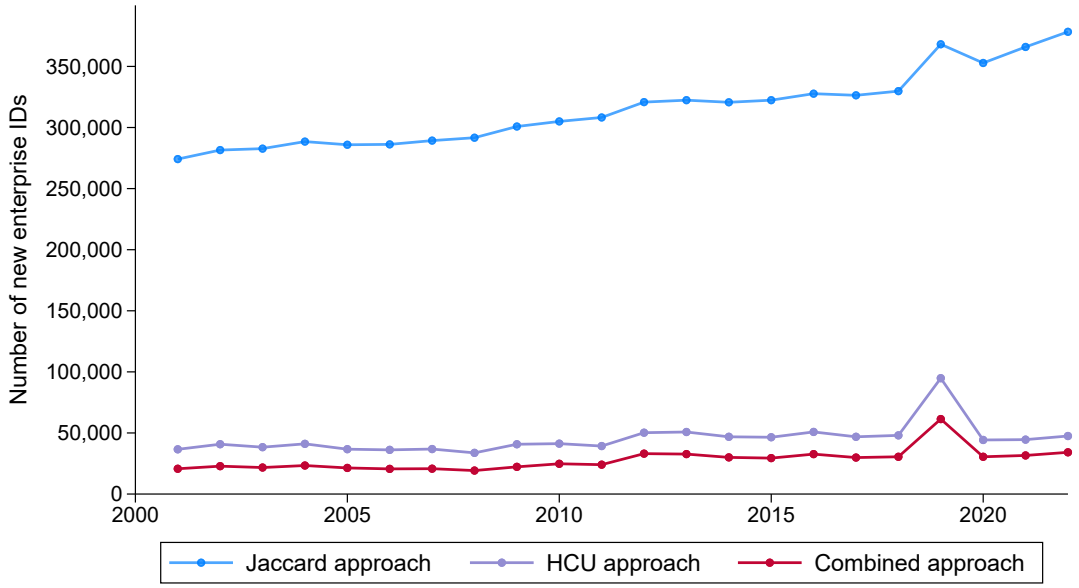


Figure A.1: New Enterprise Identifiers under Alternative Ownership-Based Linking Rules

Notes: The figure reports the annual number of newly assigned ownership-based enterprise identifiers under three alternative longitudinal linking rules: Jaccard similarity only, highest controlling unit (HCU) only, and the combined Jaccard+HCU procedure. A new identifier is assigned whenever the respective method does not classify an ownership cluster as a continuation of a cluster observed in the previous year. The sample is restricted to ownership clusters containing at least one legal unit registered in the commercial business register. Newly assigned enterprise identifiers should not be interpreted as firm entries. They may also result from ownership changes, corporate restructuring, changes in cluster composition, or the appearance of newly registered legal units. The substantially higher number of new identifiers under the Jaccard-only approach reflects the sensitivity of cluster-composition measures to changes in the underlying set of legal units, particularly among smaller ownership clusters.

At the same time, the figure illustrates the value added of incorporating Jaccard similarity as a secondary refinement criterion. Around 2019, the HCU-only approach exhibits a pronounced increase in newly assigned identifiers, indicating a rise in observed changes to highest controlling units. During this period, the combined procedure assigns noticeably fewer new identifiers than the HCU-only approach. These cases correspond to ownership clusters whose highest controlling unit changes, but whose underlying composition remains sufficiently stable to suggest continuity of the economic entity. By preserving identifiers in such cases, the combined procedure avoids excessive fragmentation of ownership-based firm histories.

Importantly, newly assigned enterprise identifiers should not be interpreted as firm entries. A new identifier merely indicates that the ownership-based clustering algorithm no longer regards the observed structure as a continuation of a previously existing ownership cluster. Such changes may arise from acquisitions, divestitures, internal reorganizations, changes in ownership shares, or the appearance of new legal units in the business register. They therefore reflect changes in observed ownership structures rather than necessarily the creation of new firms.

Taken together, the evidence suggests that the combined Jaccard+HCU procedure strikes a useful balance between stability and responsiveness. Relative to the Jaccard-only approach, it avoids excessive identifier turnover driven by minor changes in cluster composition. Relative to the HCU-only approach, it prevents some ownership changes from generating new identifiers when the broader ownership structure remains largely intact. The resulting enterprise identifiers therefore provide a more stable representation of economically meaningful firm structures over time.

A.2 Illustrative Algorithm Examples

Table A.1: Example of Ownership-Based Cluster Assignment

crefo	*b_crefo*	share	cluster_id_layer1	cluster_id_layer2
10001	99999	65%	11	1
10002	99999	70%	11	1
20001	10001	70%	11	1
30001	10002	20%	21	2
30002	10001	30%	23	1
30002	10002	30%	24	1
40001	99998	62%	12	3
40002	99998	68%	12	3
40003	99997	100%	13	4
50001	40001	75%	12	3
50001	40002	25%	14	3
50002	40001	25%	25	3
50002	40002	30%	26	3
60001	40003	70%	15	4

Notes: The table reports the ownership relations underlying Figure 5 and illustrates how ownership-based clusters are assigned across algorithmic layers. The first layer groups legal units connected through direct or indirect majority ownership. Rows highlighted in orange denote cases in which no single cluster member holds a majority stake, but aggregated ownership within the first-layer cluster exceeds 50%. The final column reports the resulting ownership-based clusters after the second layer.

Table A.2: Example of Name-Based Cluster Assignment

crefo name 1	crefo name 2	score	same cluster	cluster_id
A1	A2	85%	yes	1
A1	A3	90%	yes	1
A2	A5	90%	yes	1
A2	A3	83%	yes	1
A3	A5	81%	yes	1
B2	B3	95%	yes	2
B2	B4	85%	yes	2
B2	B5	85%	yes	2
B3	B4	80%	yes	2
B3	B5	82%	yes	2
B4	B5	95%	yes	2
C1	C2	85%	yes	3
C2	C3	100%	yes	3

Notes: The table reports the name-matching relations underlying Figure 6 and illustrates how pairwise business-name matches are assigned to cross-sectional name-based clusters. Pairwise matches with a similarity score of at least 85% define direct links. Rows highlighted in orange denote pairs with similarity below 85% that nevertheless belong to the same cluster because both units are linked indirectly through at least one above-threshold match to another unit. Singleton clusters that do not reach the threshold with any other unit are not shown.



Download ZEW Discussion Papers:

https://www.zew.de/en/publications/zew-discussion-papers

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

IMPRINT

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de

Discussion Papers are intended to make results of ZEW research promptly available to other economists in order to encourage discussion and suggestions for revisions. The authors are solely responsible for the contents which do not necessarily represent the opinion of the ZEW.