Maier, Alexander

Article

Mikrodatenverknüpfung ohne eindeutige Identifikatoren am Beispiel der Finanzdienstleistungsstatistik

WISTA - Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Maier, Alexander (2024) : Mikrodatenverknüpfung ohne eindeutige Identifikatoren am Beispiel der Finanzdienstleistungsstatistik, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 76, Iss. 2, pp. 97-106

MIKRODATENVERKNÜPFUNG OHNE EINDEUTIGE IDENTIFIKATOREN AM BEISPIEL DER FINANZDIENSTLEISTUNGSSTATISTIK

Alexander Maier

Schlüsselwörter: Einzeldatenverknüpfung – Unternehmensregister – Verwaltungsdaten – reguläre Ausdrücke – Fuzzy-Matching

ZUSAMMENFASSUNG

Für die verwaltungsdatenbasierte Finanzdienstleistungsstatistik ist die Verknüpfung von Einzeldaten des statistischen Unternehmensregisters und der Finanzaufsicht des Bundes essenziell. Datenverknüpfungen sind allerdings nicht ohne Weiteres möglich, wenn keine eindeutigen Identifikatoren vorliegen. Der Artikel beschreibt, wie in solchen Fällen dennoch erfolgreich Einzeldaten, unter Verwendung regulärer Ausdrücke sowie des sogenannten Fuzzy-Matchings, maschinell verknüpft werden können. Erstmals wurde dieses neu entwickelte Verfahren in der Finanzdienstleistungsstatistik im Zuge der Umsetzung der EBS-Verordnung implementiert.

Keywords: *microdata linkage – business register – administrative data – regular expressions – fuzzy matching*

ABSTRACT

Linking microdata from the statistical business register and from federal financial supervisory authorities is essential for the statistics on financial services, which is based on administrative data. However, linkages are usually not possible without unique identifiers. This paper describes how microdata can nevertheless be linked automatically in such cases by using regular expressions and fuzzy matching. This newly developed method was applied for the first time in the statistics on financial services during the implementation of the EBS Regulation.

Alexander Maier

ist Ökonom und seit Oktober 2021 wissenschaftlicher Mitarbeiter im Referat „Struktur des Handels und der Dienstleistungen“ des Statistischen Bundesamtes. Sein Aufgabenschwerpunkt lag darin, die EBS-Verordnung in der Finanzdienstleistungsstatistik umzusetzen. Hierfür hat er Schätzverfahren entwickelt, die auch in anderen Unternehmensstrukturstatistiken zum Einsatz kommen.

1

Einleitung

Mit dem Berichtsjahr 2021 wurde in den Unternehmensstatistiken erfolgreich die European-Business-Statistics (EBS)-Verordnung umgesetzt, wodurch die Wirtschaftsstrukturen in Deutschland deutlich umfassender abgebildet werden. Für die Finanzdienstleistungsstatistik bedeutete dies die erstmals vollständige Abdeckung des Wirtschaftsabschnitts K (Erbringung von Finanz- und Versicherungsdienstleistungen) der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (Allafi und andere, 2022; Statistisches Bundesamt, 2008).[1] Als weitere Neuerung wurde der bestehende Merkmalskranz vonseiten der Europäischen Union (EU) teilweise angepasst und erweitert, wodurch neue Schätzverfahren nötig geworden sind.

Zugleich verarbeitete die Finanzdienstleistungsstatistik erstmals Einzel- statt aggregierter Daten, um die Darstellungseinheit Unternehmen bilden zu können, die in der deutschen amtlichen Statistik aus der Rechtlichen Einheit abgeleitet wird (Beck und andere, 2020).[2] Im Zuge dessen wurde ebenfalls erstmals die Grundgesamtheit der Finanzdienstleistungsstatistik auf Basis des statistischen Unternehmensregisters (URS) für Rechtliche Einheiten gebildet. Damit erhöht sich zugleich die Kohärenz der Finanzdienstleistungsstatistik zur Unternehmensdemografie.

Bislang wurde die Finanzdienstleistungsstatistik rein auf Basis von Verwaltungsdaten erstellt. Die vielfältigen Änderungen erforderten jedoch einen kosteneffizienten und belastungsarmen Methodenmix aus einer Primärerhebung und der hauptsächlichen Nutzung von Register- beziehungsweise Verwaltungsdaten (Allafi und andere, 2022). Vor diesem Hintergrund erfolgte im Jahr 2022 die Anpassung des § 3b Verwaltungsdatenverwendungsgesetz in Einklang mit § 5a Absatz 1 Bundesstatistikgesetz[3]. Mithilfe der erweiterten Verwaltungsdatennutzung können weitere Erhebungen vermieden (Once-Only-Prinzip[4]), Register ergänzt und gepflegt sowie die Qualität der Daten der amtlichen Statistik gesichert werden (Bens/Schukraft, 2018).

Um auf Basis der URS-Grundgesamtheit überhaupt Verwaltungsdaten nutzen zu können, sind wiederum Datenverknüpfungen nötig. Die statistikrechtliche Grundlage hierfür existiert grundsätzlich mit dem 1990 eingefügten § 13a Bundesstatistikgesetz. Seit seiner Neufassung im Jahr 2005 enthält dieser explizit die Erlaubnis, Daten nach dem Verwaltungsdatenverwendungsgesetz zu verknüpfen. Die verbesserte Nutzbarkeit von Register- und Verwaltungsdaten erfüllt mehrere der vom Statistischen Beirat[5] (2010) geforderten Ziele hinsichtlich der Weiterentwicklung der amtlichen Statistik.

In dieselbe Richtung gingen EU-Projekte wie das „Microdata linking of Structural Business Statistics and other business statistics“ (MDL). Hierfür wurden Einzeldaten aus zwölf verschiedenen Unternehmensstrukturstatistiken (ausgenommen die Finanzdienstleistungsstatistik), der Unternehmensdemografie, der Statistik über Auslandsunternehmenseinheiten und der Erhebung über die Nutzung von Informations- und Kommunikationstechnologien über verschiedene Berichtsjahre hinweg verknüpft (Jung/Käuser, 2016).

Dabei stellt bereits die Verknüpfung von Mikrodaten innerhalb der amtlichen Statistik diese vor eine Vielzahl von Herausforderungen. Dies ist vor allem bedingt durch unterschiedliche Erhebungsdesigns, welche die mögliche Schnittmenge an identischen Einheiten reduzieren, und weniger durch das Fehlen von eindeutigen Identifikatoren (Jung/Käuser, 2016). Ein weiterer Grund kann der sich ändernde Schwerpunkt wirtschaftlicher Aktivitäten einer Einheit sein, der für die nach Wirtschaftsabschnitten abgegrenzten Unternehmensstrukturstatistiken relevant ist, nicht jedoch für andere Statistiken oder

1 Dabei wurde das Statistische Bundesamt im Rahmen des EU-Grants „SBS: Support to set up the production of variables for the new Section K NACE codes“ 2021 und 2022 finanziell unterstützt.

2 Gemäß EU-Definition entspricht ein Unternehmen der kleinsten Kombination Rechtlicher Einheiten, die eine organisatorische Einheit zur Erzeugung von Waren und Dienstleistungen bildet. In anderen Unternehmensstrukturstatistiken wird bereits seit dem Berichtsjahr 2018 der EU-Unternehmensbegriff verwendet.

3 Nach § 5a Absatz 1 Bundesstatistikgesetz ist seit 2016 pflichtgemäß vor der Einführung neuer oder der Änderung bestehender Bundesstatistiken zu prüfen, ob in der öffentlichen Verwaltung oder bei ähnlichen Stellen bereits geeignete Daten vorliegen.

4 Dessen Ziel ist, dass Personen und Unternehmen bestimmte Informationen nur noch einmal der Verwaltung mitteilen müssen und diese für die Weitergabe der Daten an relevante Stellen sorgt.

5 Der Statistische Beirat berät gemäß § 4 Bundesstatistikgesetz das Statistische Bundesamt in Fachfragen und vertritt die Belange der Nutzerinnen und Nutzer der Bundesstatistik.

Verwaltungsdaten. Bei der Verknüpfung können zudem Inkohärenzen auftreten.[6]

Im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz wird im Projekt „Methodische und analytische Stärkung in aktuellen Fragen der Außenhandels- und ausländischen Investitionspolitik“ bis Mai 2024 an Lösungen für derartige Probleme bei der Verknüpfung von Daten der Außenhandels- und Unternehmensstatistiken gearbeitet (Kruse und andere, 2021).

Allerdings führen etliche Behörden Datenbanken mit eigenen Identifikatoren, die nicht unbedingt aufeinander abgestimmt sind (McKinsey, 2017). Bei der Finanzdienstleistungsstatistik fehlen dadurch häufig eindeutige Identifikatoren, anhand derer Einzeldaten verknüpft werden könnten. Zudem hat hier bislang keine umfassende primärstatistische Erhebung stattgefunden, auf deren Basis das URS in der Vergangenheit hätte gepflegt werden können.

Dieser Artikel befasst sich vorrangig mit der Mikrodatenverknüpfung ohne eindeutige Identifikatoren und beginnt in Kapitel 2 mit einem Überblick über die zu verknüpfenden Datensätze. Anschließend geht Kapitel 3 genauer auf die speziellen Herausforderungen der Datenverknüpfung in der Finanzdienstleistungsstatistik ein und stellt das Verfahren zur Aufbereitung und Verknüpfung dar. Abschließend wird in Kapitel 4 ein Fazit gezogen und ein Ausblick gegeben.

2

Datenbasis

Um EU-Lieferverpflichtungen zu erfüllen und Doppelerhebungen zu vermeiden, erhält das Statistische Bundesamt Einzeldaten von der Finanzaufsicht des Bundes[7]. Die Daten stammen überwiegend aus dem externen Rechnungswesen der zu beaufsichtigenden Rechtlichen Einheiten; je nach Branche sind sie aufgrund spezieller Verordnungen[8] an die jeweils zuständige Aufsichtsbehörde zu übermitteln. Seit der Novellierung des Verwaltungsdatenverwendungsgesetzes 2022 trifft dies für folgende Wirtschaftszweige der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008; Statistisches Bundesamt, 2008) zu:

- Zentralbanken und Kreditinstitute (64.1),
- Sonstige Finanzierungsinstitutionen (64.9),
- Versicherungen, Rückversicherungen und Pensionskassen (ohne Sozialversicherung) (65) und
- Mit Finanzdienstleistungen verbundene Tätigkeiten (66.1).[9]

Um Datenverknüpfungen zu ermöglichen, erhält das Statistische Bundesamt darüber hinaus die Einheitenbezeichnungen sowie Handelsregister- (HR-) und Umsatzsteuer- (UST-) Identifikationsnummern (ID), den Legal Entity Identifier (LEI)[10] und weitere Angaben, beispielsweise die Anschrift.

Mit dem Ziel, die restlichen Lücken im Wirtschaftsabschnitt K zu schließen, wird derzeit bei den Wirtschaftsgruppen „Beteiligungsgesellschaften (64.2)“[11] und „Treuhand- und sonstige Fonds und ähnliche Finanzinstitutionen (64.3)“ von Unternehmensregisterdaten Gebrauch gemacht. Im Unterschied zur Finanzaufsicht des Bundes bezieht das URS im Kern seine Angaben zu Einheiten mittelbar von der Bundesagentur für Arbeit, den Finanzverwaltungen der Länder und verschiedenen Registern, darunter dem Handelsregister. Es verfügt für alle Einheiten über eine eindeutige URS-ID sowie für etliche auch über die HR- und UST-ID, derzeit aber nicht über den im Finanzsektor geläufigen LEI.

Die Wirtschaftsgruppe 66.2 „Mit Versicherungsdienstleistungen und Pensionskassen verbundene Tätigkeiten“ ist die einzige aus dem Wirtschaftsabschnitt K, für die keine in der öffentlichen Verwaltung vorliegenden, geeigneten Daten ausfindig gemacht werden konnten.

6 Zum Beispiel enthält das URS nur Meldungen zum steuerbaren Umsatz. Angaben zum nicht steuerbaren Umsatz, der im Finanzdienstleistungsbereich durchaus vorkommt, fehlen hingegen (siehe § 4 Umsatzsteuergesetz).

7 Sie besteht aus der Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin) und der Deutschen Bundesbank. Die Datenlieferungen sind in den §§ 3a und 3b Verwaltungsdatenverwendungsgesetz geregelt.

8 Kreditinstituts-Rechnungslegungsverordnung, Pensionsfonds-Aufsichtsverordnung oder Versicherungsberichterstattungs-Verordnung.

9 Die Wirtschaftszweige 64.9 und 66.1 sind erst im Zuge der Novellierung des Verwaltungsdatenverwendungsgesetzes 2022 hinzugekommen. Bereits vorher wurden die Spezialkreditinstitute (64.92) berücksichtigt.

10 Der LEI ist eine eindeutige internationale Identifikationsnummer für finanzielle Kapitalgesellschaften.

11 Ohne Verwaltungs- und Managementfunktion.

Übersicht 1
Wirtschaftsabschnitt K[1] und geeignete Datenquellen

WZ-2008-Code	Bezeichnung	Datenquelle
64	Erbringung von Finanzdienstleistungen	
64.1	Zentralbanken und Kreditinstitute	Deutsche Bundesbank (Verwaltungsdaten)
neu: 64.2	Beteiligungsgesellschaften	Statistisches Unternehmensregister
neu: 64.3	Treuhand- und sonstige Fonds und ähnliche Finanzinstitutionen	Statistisches Unternehmensregister
neu: 64.9	Sonstige Finanzierungsinstitutionen	Deutsche Bundesbank (Verwaltungsdaten)
65	Versicherungen, Rückversicherungen und Pensionskassen (ohne Sozialversicherung)	Bundesanstalt für Finanzdienstleistungsaufsicht (Verwaltungsdaten)
65.1	Versicherungen	
65.2	Rückversicherungen	
65.3	Pensionskassen und Pensionsfonds	
66	Mit Finanz- und Versicherungsdienstleistungen verbundene Tätigkeiten	
neu: 66.1	Mit Finanzdienstleistungen verbundene Tätigkeiten	Deutsche Bundesbank (Verwaltungsdaten)
neu: 66.2	Mit Versicherungsdienstleistungen und Pensionskassen verbundene Tätigkeiten	Erhebungsdaten (Strukturstatistik im Handels- und Dienstleistungsbereich)
neu: 66.3	Fondsmanagement	Deutsche Bundesbank

1 Der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008).

Um Synergieeffekte zu nutzen, wurde diese in die neue Strukturerhebung im Handels- und Dienstleistungsbereich integriert. Das bedeutet, dass der deutlich ausgeweitete Erfassungsbereich nahezu vollständig von der verwaltungs- und registerbasierten Finanzdienstleistungsstatistik aufwandsarm abgedeckt wird (Allafi und andere, 2022). ↘ Übersicht 1

Für etwa jede dritte Rechtliche Einheit aus der auf dem Unternehmensregister basierenden Grundgesamtheit der Finanzdienstleistungsstatistik (ohne Wirtschaftsgruppe 66.2) ist es somit möglich, Ergebnisse für die an Eurostat[12] zu liefernden Merkmale[13] abzuleiten. Allerdings kann bezüglich der URS-Grundgesamtheit eine Über- oder Untererfassung vorliegen, vor allem durch unternehmensdemografische Ereignisse wie Übernahmen, Schließungen oder Neugründungen. Zudem sind darin keine Einheiten enthalten, die nicht über mindestens eine abhängig beschäftigte Person verfügen.

Darüber hinaus ist die Grundgesamtheit der zu beaufsichtigenden Institute durch die Finanzaufsicht des Bundes nicht mit dem Wirtschaftsabschnitt K deckungsgleich: Zum einen werden auch Einheiten außerhalb dieses Abschnitts beaufsichtigt, zum anderen werden nicht alle Einheiten des Abschnitts beaufsichtigt, sodass die Schnittmenge nicht den vollständigen Verwaltungsdatensatz umfasst.

3 Herausforderungen der Mikrodatenverknüpfung in der Finanzdienstleistungsstatistik

Register- und Verwaltungsdaten werden nicht primär für statistische Zwecke erhoben und genügen daher nicht immer den Anforderungen, die die amtliche Statistik an ihre Daten stellt. Die unterschiedlich geführten Datenbanken, die verschiedenen Abgrenzungen sowie die nötige Verknüpfung auf Ebene der Einzeldaten führen dazu, dass sich die Nutzung solcher Daten für die Finanzdienstleistungsstatistik entsprechend herausfordernd darstellt.

12 Eurostat ist das Statistische Amt der EU.

13 Diese Merkmale gibt die Durchführungsverordnung (EU) 2020/1197 vor.

3.1 Schwierigkeiten bei der Datenaufbereitung

Die verschiedenen Datenquellen führen in der Praxis zu unterschiedlichen Dateiformaten und allgemein nicht harmonisierten Inputdateien. Dafür gehen die Erfassungsschemata der Finanzaufsichtsbehörden in der Regel über die Angaben in den Jahresabschlüssen der Einheiten hinaus. Oft genug liegen auch keine öffentlich zugänglichen Jahresabschlüsse vor, da nicht alle Rechtlichen Einheiten aus dem Wirtschaftsabschnitt K dazu verpflichtet sind, diese offenzulegen.[14]

Bei den erhobenen Daten der Deutschen Bundesbank konnte das Statistische Bundesamt in Zusammenarbeit mit den jeweils zuständigen Fachbereichen den Aufbereitungsaufwand durch heterogene Dateiformate minimieren. Daher war die für diesen Zweck angepasste IT-Anwendung[15] in der Lage, die von der Deutschen Bundesbank gelieferten Inputdateien aufwandsarm maschinell einzulesen und aufzubereiten.

Insbesondere die von der Bundesanstalt für Finanzdienstleistungsaufsicht gelieferte Inputdatei erforderte jedoch für das Berichtsjahr 2021 zunächst aufwendige Anpassungen der Aufbereitungsprozedur. Grund dafür war, dass je nach Wirtschaftszweig und Einheit unterschiedliche und unterschiedlich viele Informationen vorliegen konnten, die für die weitere Bearbeitung zunächst in einer Zeile je Einheit verdichtet werden mussten.

Die Daten der Finanzaufsicht des Bundes wurden anschließend zusammengeführt. In der Regel waren die für die Lieferung an Eurostat erforderlichen Merkmale noch nicht in den aufbereiteten und plausibilisierten Verwaltungsdaten enthalten.[16]

3.2 Die Problematik der Einzeldatenverknüpfung

Wesentliche Voraussetzung für die einfache Verknüpfung von Mikrodaten ist das Vorliegen mindestens eines gemeinsamen und eindeutigen Identifikators. Allerdings führt die Finanzaufsicht des Bundes Datenbanken nicht primär für statistische Zwecke und verwendet zum Teil andere Identifikatoren.

In der Finanzdienstleistungsstatistik eignen sich zum Beispiel grundsätzlich die UST-ID, die HR-ID oder die von der jeweiligen Verwaltungsdatenquelle vergebene Aufsichts-ID sowie der LEI. Jedoch verfügt das URS weder über die in den Verwaltungsdaten vorkommende Aufsichts-ID noch über den LEI; es verwendet eine eigene URS-ID. Des Weiteren liegen nur für eine Teilmenge von Einzeldaten beider Datensätze die erwähnten Identifikatoren vor. ↘ Übersicht 2

Die HR-ID ist zudem erst in Kombination mit der Art des Handelsregisters[17] und dem Eintragungsort in das jeweilige Register (Registersitz) eindeutig. Jedoch gibt

14 Solche Befreiungsregelungen sind in § 264 Absatz 3 oder § 264b Handelsgesetzbuch (HGB) aufgeführt.

15 Die im Folgenden skizzierte, fachbereichsinterne IT-Anwendung basiert auf SAS, einer Software für Statistikanalyse und Datenauswertung.

16 Für die Ableitung der Merkmale ist es meist nötig, Zwischenmerkmale zu berechnen oder top down zu schätzen.

17 In Deutschland gibt es zum einen zwei Abteilungen (A und B) des Handelsregisters, zum anderen noch weitere Register, wie das Genossenschafts-, das Partnerschafts- oder das Vereinsregister. Auch die Letztgenannten werden in diesem Artikel unter dem Begriff Handelsregister subsumiert.

Übersicht 2

Identifikatoren bei relevanten Register- und Verwaltungsdaten 2021

Datenquelle	Institutsname	Aufsichts-Identifikationsnummer	Umsatzsteuer-Identifikationsnummer	Handelsregister-Angaben[1]	Legal Entity Identifier (LEI)[2]	Weitere Angaben[3]
Bundesanstalt für Finanzdienstleistungsaufsicht (Verwaltungsdaten)	ja	ja	nein	teilweise	teilweise	ja
Deutsche Bundesbank (Verwaltungsdaten)	ja	ja	teilweise	teilweise	teilweise	ja
Statistisches Unternehmensregister	ja	nein	teilweise	teilweise	nein	ja

1 Diese Angaben umfassen die Registerart, die Registernummer sowie den Registersitz.
2 Der LEI ist eine eindeutige internationale Identifikationsnummer für finanzielle Kapitalgesellschaften.
3 Die weiteren Angaben umfassen unter anderem die Anschrift und Rechtsform.

es sowohl im URS, dessen Quelle unmittelbar das jeweilige Register ist, als auch in den Verwaltungsdaten der Finanzaufsicht des Bundes bereits bei den Registersitzen unterschiedliche Schreibweisen derselben Orte.[18]

Dieselbe Problematik tritt in noch größerem Ausmaß bei den Institutsnamen und den Adressen auf, bei denen das Unternehmensregister die Angaben ebenfalls direkt aus dem jeweiligen Handelsregister bezieht. Dies liegt an Unterschieden in der Verwendung von Abkürzungen und Schreibweisen sowie bezüglich der Satz-, Schrift- und Leerzeichen. Darum sind diese alphanumerischen[19] Identifikatoren grundsätzlich nicht so gut geeignet wie rein numerische. Bei den Anschriften kommt noch hinzu, dass diese weniger zuverlässig sind, da unter einer Adresse viele Rechtliche Einheiten gemeldet sein können. Somit entfallen drei weitere infrage kommende Identifikatoren mangels Eindeutigkeit für eine einfache Verknüpfung.

Zu beachten ist zudem, dass durch die unterschiedliche Abgrenzung der Grundgesamtheiten von Unternehmensregister und Verwaltungsdaten (siehe Kapitel 2) beaufsichtigte Einheiten in einem Wirtschaftszweig verortet sein können, der außerhalb des Wirtschaftsabschnitts K (ohne Wirtschaftsgruppe 66.2) liegt.[20] In diesem Fall dienen die Verwaltungsdaten dazu, den Wirtschaftszweig zu überprüfen. So konnten mit den vorliegenden Identifikatoren zunächst lediglich rund 60 % der Verwaltungsdaten korrekt mit der URS-Grundgesamtheit verknüpft werden,[21] und zwar überwiegend Kreditinstitute.

3.3 Auswege aus der Verknüpfungsproblematik

Um in der Finanzdienstleistungsstatistik eine gewisse Qualität vor allem in den Wirtschaftsgruppen 64.9, 66.1 und 66.3 herzustellen, ist dementsprechend vor allem das Potenzial der nicht eindeutigen, aber grundsätzlich zuverlässigen Identifikatoren auszuschöpfen, also der Einheitenbezeichnungen sowie der Handelsregister-Angaben. Hierfür bieten sich zwei Verfahren an (Schneider, 2019; Sloan/Lafler, 2022).

Eine Möglichkeit ist, die alphanumerischen Identifikatoren in den vorliegenden Verwaltungs- und URS-Daten zu harmonisieren und so eine höhere Ähnlichkeit oder sogar Eindeutigkeit herzustellen. Hierfür liegt es nahe, zunächst gewisse Muster im Datensatz zu identifizieren. Dies können Ausdrücke sein, die sich gehäuft in ähnlicher Form in beiden Datensätzen finden.

Zum Beispiel befinden sich unter den Einheiten viele Aktiengesellschaften, Gesellschaften mit beschränkter Haftung und Versicherungsvereine auf Gegenseitigkeit oder Anstalten des öffentlichen Rechts. In der Regel werden dabei die Rechtsformen als Bestandteil der Institutsnamen uneinheitlich aufgeführt. Werden diese identifiziert und die verschiedenen Schreibweisen durch gängige Ausdrücke wie AG, GmbH, VVaG beziehungsweise AöR in beiden Datensätzen gleichermaßen ersetzt, so ähneln sich die Bezeichnungen sehr viel mehr oder sind im Idealfall sogar identisch.

Dies trifft auch auf Branchenbezeichnungen zu, wie etwa Kranken-, Lebensversicherung, Pensionskasse oder Kapitalverwaltungsgesellschaft. Sie können ebenfalls mittels gängiger Abkürzungen harmonisiert werden, zum Beispiel KV, LV, PK oder KVG. Zudem können Bezüge zu Inhabern, Gründungsdaten, Verweise auf Standorte und Sonderzeichen, die Teil der Einheitenbezeichnungen sind, vereinheitlicht beziehungsweise entfernt werden. Wesentlich weniger Abweichungen liegen bei den Registersitzen vor.

Die manuelle Identifizierung aller gleichbedeutenden, aber abweichenden Ausdrücke wäre allerdings sehr zeitaufwendig und bliebe wahrscheinlich auch unvollständig. Abhilfe schafft hier die Verwendung regulärer Ausdrücke der freien Programmiersprache Perl, sogenannte perl regular expressions (PRX). Mit ihnen werden ähnliche Ausdrücke beziehungsweise Muster in den Datensätzen maschinell gesucht und zugleich die identifizierten durch einheitliche Ausdrücke substituiert.[22]

18 Zum Beispiel konnte der Autor im untersuchten Datensatz drei verschiedene Schreibweisen für das Amtsgericht Charlottenburg in Berlin identifizieren.

19 Alphanumerische Merkmale weisen neben Ziffern und gegebenenfalls Operations- beziehungsweise Sonderzeichen mindestens einen Buchstaben eines Alphabets auf.

20 Zum Beispiel Bausparkassen und ähnliche Einrichtungen oder Holdings.

21 Die Verknüpfung erfolgt im Fachbereich mittels der Datenbanksprache SQL in SAS.

22 Für technische Details siehe beispielsweise Wall/Schwartz (1991) sowie Windham (2014).

Anschließend werden die Mikrodaten aus dem URS- und dem Verwaltungsdatensatz in einem mehrstufigen Verfahren verknüpft. Im ersten Schritt erfolgt dies über die eindeutige UST-ID. Im zweiten Schritt werden beide Einzeldatensätze über die harmonisierten Identifikatoren, Handelsregister-Art, -ID und den Registersitz sowie die Einheitenbezeichnungen, kombiniert.

Eine weitere Möglichkeit besteht darin, die Suche nach Ähnlichkeiten in den (nicht) harmonisierten Datensätzen grundsätzlich unscharf durchzuführen, mit der sogenannten Fuzzy-Suche. Diese basiert auf der von Zadeh (1973) formulierten Fuzzylogik und besteht darin, in zwei Zeichenfolgen (Strings) nach Ähnlichkeiten zu suchen. Dabei wird als Ähnlichkeitsmaß zum Beispiel die von Levenshtein (1966) entwickelte Levenshtein-Distanz genutzt. Diese gibt die Zahl der Operationen an, die benötigt wird, um von dem Vergleichs- auf den Basisstring zu gelangen. Auf dieser Grundlage kann die Grundgesamtheit mit den Einheiten verknüpft werden, für die die Ähnlichkeit am größten ausfällt. Diese Methode ist als Fuzzy-Matching bekannt.

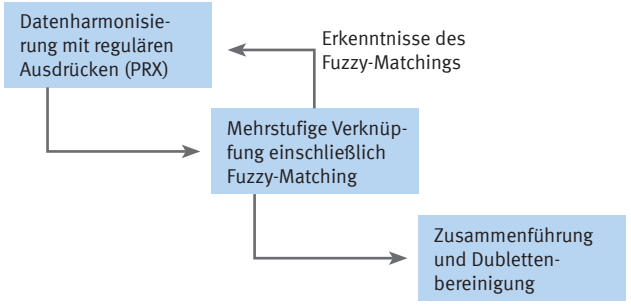
Speziell bei Verwendung harmonisierter Datensätze – und vor allem, wenn diese über die Adressen verknüpft werden – kommen allerdings sehr viele Mehrfachverknüpfungen zustande. Zwar kann eine Obergrenze für die Levenshtein-Distanz eingeführt werden, um die Anzahl der Mehrfachverknüpfungen zu reduzieren und die Genauigkeit der Verknüpfungen zu erhöhen. Jeder Wert wäre jedoch zwangsläufig (anfangs) willkürlich und könnte dazu führen, dass die korrekte Verknüpfung abhandenkommt.

Daher eignet sich diese Methode nicht unbedingt für ein automatisiertes Verfahren, sondern als Ergänzung zu den PRX. Die manuell als korrekt identifizierten Verknüpfungen infolge des Fuzzy-Matchings können dann über eine Anpassung des beschriebenen Harmonisierungsprozesses bei einem weiteren Durchlauf maschinell verknüpft werden.

Letztlich können allerdings aus dem mehrstufigen Verknüpfungsprozess, in dem zuerst die eindeutigen und dann die nicht eindeutigen Identifikatoren herangezogen werden, Mehrfachverknüpfungen erfolgen. Die Duplikate werden daher maschinell über die URS-ID unter Zuhilfenahme der Aufsichts-ID und eines Abgleichs der Rechtsformen eliminiert.

Um die vorhandenen Verwaltungsdaten optimal zu nutzen, werden beide Lösungsansätze in der Finanzdienstleistungsstatistik seit dem Berichtsjahr 2021 eingesetzt. Zunächst werden die aufbereiteten Verwaltungsdaten und die URS-Grundgesamtheit hinsichtlich der nicht eindeutigen Identifikatoren vereinheitlicht. Anschließend wird mithilfe der nutzbaren Identifikatoren die Datenverknüpfung durchgeführt. Ergänzend wird versucht, mit Fuzzy-Matching weitere, verknüpfbare Einheiten aufzuspüren. Die daraus gewonnenen Erkenntnisse fließen in den automatisierten Harmonisierungsprozess ein. Schlussendlich wird der kombinierte Datensatz aus URS-Grundgesamtheit und Verwaltungsdaten maschinell um Dubletten bereinigt. ↘ Grafik 1

Grafik 1
Das neue Verfahren zur Mikrodatenverknüpfung in der Finanzdienstleistungsstatistik



Mithilfe des neuen Verfahrens konnte die Quote der korrekt verknüpften Verwaltungsdaten erfolgreich von etwa 60 % auf rund 75 % gesteigert werden. Da die zusätzlich verknüpften Einheiten vor allem aus dem Bereich der Versicherungen und Pensionskassen sowie den sonstigen Finanzdienstleistern stammen, konnte hierdurch die Finanzdienstleistungsstatistik in diesen Teilbereichen qualitativ erheblich verbessert werden.

Prozessoptimierungen sind sowohl bei der IT-Anwendung als auch durch die Pflege des URS möglich und wurden bereits für das Berichtsjahr 2022 teilweise umgesetzt. Hierdurch wird sich voraussichtlich die Quote der verknüpften Verwaltungsdaten, bei gleichbleibenden wirtschaftlichen Schwerpunkten der Einheiten, künftig weiter erhöhen. Durch die unterschiedlichen Abgrenzungen hinsichtlich der Wirtschaftszweige werden sich jedoch auch mit einem weiterentwickelten Verfahren nicht alle Verwaltungsdaten verknüpfen lassen.

4

Fazit

Seit dem Berichtsjahr 2021 werden erstmals für den gesamten Wirtschaftsabschnitt K Ergebnisse gemäß der EBS-Verordnung und der EU-Unternehmensdefinition erstellt. Durch die gemäß Once-Only-Prinzip verstärkte Nutzung von Verwaltungsdaten konnten die zusätzlichen Belastungen für die Wirtschaft minimiert werden. Daneben führten die gewonnenen Erkenntnisse über den Umsatz der Einheiten dazu, die entsprechenden Angaben im URS zu aktualisieren.

Dieses kosteneffiziente und vielversprechende Vorgehen ist seit den Novellierungen des Bundesstatistikgesetzes in den Jahren 2005 und 2016 rechtlich zulässig sowie geboten. Zudem kann in Kombination mit dem Verwaltungsdatenverwendungsgesetz die Verwaltungsdatennutzung in Zukunft vergleichsweise flexibel erweitert werden.

In der Praxis bleibt es durch unterschiedlich geführte Datenbestände mit uneinheitlich verwendeten Identifikatoren problematisch, Einzeldaten der amtlichen Statistik mit Verwaltungsdaten aus anderen Quellen zu verknüpfen. Durch die Harmonisierung nicht eindeutiger Identifikatoren sowie Fuzzy-Matching lassen sich jedoch zum einen mehr Identifikatoren nutzen und können zum anderen die Verknüpfungen weiterhin maschinell hergestellt werden. Das neue Verfahren zur Mikrodatenverknüpfung in der Finanzdienstleistungsstatistik ist somit auch für andere amtliche Statistiken und Register relevant.

In nicht allzu ferner Zukunft wird sich die Problematik der Mikrodatenverknüpfung durch das beim Statistischen Bundesamt als Registerbehörde errichtete und betriebene Basisregister für Unternehmen vereinfachen. In diesem Verwaltungsregister werden Identifikatoren, darunter der LEI, aller in der deutschen Verwaltung geführten Unternehmen zusammengeführt. Insbesondere wird eine bundeseinheitliche Wirtschaftsnummer je (Rechtlicher) Einheit eingeführt, wofür das Unternehmensbasisdatenregistergesetz von 2021 die Grundlage geschaffen hat.

LITERATURVERZEICHNIS

Allafi, Sabine/Lohn, Alexandra/Nölting, Christopher/Maier, Alexander. *Die neue Strukturstatistik im Handels- und Dienstleistungsbereich*. In: WISTA Wirtschaft und Statistik. Ausgabe 5/2022, Seite 22 ff.

Beck, Martin/Baumgärtner, Luisa/Bürk, Katja-Verena/Redecker, Matthias. *Einführung des EU-Unternehmensbegriffs: Konzept und Umsetzung*. In: WISTA Wirtschaft und Statistik. Ausgabe 3/2020, Seite 35 ff.

Bens, Arno/Schukraft, Stefan. *Registermodernisierung und Verwaltungsdatennutzung in der amtlichen Statistik*. In: WISTA Wirtschaft und Statistik. Ausgabe 4/2018, Seite 11 ff.

Jung, Sandra/Käuser, Stefanie. *Herausforderungen und Potenziale der Einzeldatenverknüpfung in der Unternehmensstatistik*. In: WISTA Wirtschaft und Statistik. Ausgabe 2/2016, Seite 95 ff.

Kruse, Hendrik W./Meyerhoff, Annette/Erbe, Anette. *Neue Methoden zur Mikrodatenverknüpfung von Außenhandels- und Unternehmensstatistiken*. In: WISTA Wirtschaft und Statistik. Ausgabe 5/2021, Seite 53 ff.

Levenshtein, Vladimir I. *Binary codes capable of correcting deletions, insertions, and reversals*. In: Soviet Physics Doklady. Ausgabe 10/1966, Seite 707 ff.

McKinsey. *Mehr Leistung für Bürger und Unternehmen: Verwaltung digitalisieren. Register modernisieren*. Gutachten im Auftrag des Nationalen Normenkontrollrats. 2017. [Zugriff am 6. März 2024]. Verfügbar unter: www.normenkontrollrat.bund.de

Schneider, Volker. *Digitalisierung in der amtlichen Statistik – Nutzung von Verwaltungsdaten*. In: Statistisches Monatsheft Baden-Württemberg. Ausgabe 5/2019, Seite 28 ff. [Zugriff am 6. März 2024]. Verfügbar unter: www.statistischebibliothek.de

Sloan, Stephen/Lafler, Kirk P. *A Quick Look at Fuzzy Matching Programming Techniques Using SAS Software*. In: PharmaSUG 2022 Conference Paper AP-030. 2022. [Zugriff am 6. März 2024]. Verfügbar unter: www.lexjansen.com

Statistischer Beirat. *Eckpunkte zur Weiterentwicklung der amtlichen Statistik in der 17. Legislaturperiode*. 2010. [Zugriff am 6. März 2024]. Verfügbar unter: https://bdi.eu

Statistisches Bundesamt. *Klassifikation der Wirtschaftszweige, Ausgabe 2008*. Wiesbaden 2008.

Wall, Larry/Schwartz, Randal L. *Programming Perl*. Erste Auflage. Sebastopol 1991.

Windham, K. Matthew. *Introduction to Regular Expressions in SAS*. Cary 2014.

Zadeh, Lotfi A. *Outline of a New Approach to the Analysis of Complex Systems and Decision Processes*. In: IEEE Transactions on Systems, Man, and Cybernetics. Ausgabe SMC-3/1973, Seite 28 ff.

RECHTSGRUNDLAGEN

Durchführungsverordnung (EU) 2020/1197 der Kommission vom 30. Juli 2020 zur Festlegung technischer Spezifikationen und Einzelheiten nach der Verordnung (EU) 2019/2152 des Europäischen Parlaments und des Rates über europäische Unternehmensstatistiken, zur Aufhebung von zehn Rechtsakten im Bereich Unternehmensstatistiken (Amtsblatt der EU Nr. L 271, Seite 1).

Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) in der Fassung der Bekanntmachung vom 20. Oktober 2016 (BGBl. I Seite 2394), das zuletzt durch Artikel 2 des Gesetzes vom 20. Dezember 2022 (BGBl. I Seite 2727) geändert worden ist.

Gesetz über die Verwendung von Verwaltungsdaten für Zwecke der Wirtschaftsstatistiken (Verwaltungsdatenverwendungsgesetz – VwDVG) vom 4. November 2010 (BGBl. I Seite 1480), das zuletzt durch Artikel 2 des Gesetzes vom 20. Dezember 2022 (BGBl. I Seite 2727) geändert worden ist.

Gesetz zur Errichtung und Führung eines Registers über Unternehmensbasisdaten und zur Einführung einer bundeseinheitlichen Wirtschaftsnummer für Unternehmen (Unternehmensbasisdatenregistergesetz – UBRegG) vom 9. Juli 2021 (BGBl. I Seite 2506), das zuletzt durch Artikel 1 des Gesetzes vom 22. Dezember 2023 (BGBl. I Nr. 404) geändert worden ist.

Handelsgesetzbuch in der im Bundesgesetzblatt Teil III, Gliederungsnummer 4100-1, veröffentlichten bereinigten Fassung, das zuletzt durch Artikel 34 Absatz 1 des Gesetzes vom 22. Dezember 2023 (BGBl. I Nr. 411) geändert worden ist.

Umsatzsteuergesetz (UStG) in der Fassung der Bekanntmachung vom 21. Februar 2005 (BGBl. I Seite 386), das zuletzt durch Artikel 18 des Gesetzes vom 11. Dezember 2023 (BGBl. I Nr. 354) geändert worden ist.

Verordnung betreffend die Aufsicht über Pensionsfonds und über die Durchführung reiner Beitragszusagen in der betrieblichen Altersversorgung (Pensionsfonds-Aufsichtsverordnung – PFAV) vom 18. April 2016 (BGBl. I Seite 842), die zuletzt durch Artikel 2 der Verordnung vom 22. April 2021 (BGBl. I Seite 842) geändert worden ist.

Verordnung (EU) 2019/2152 des Europäischen Parlaments und des Rates vom 27. November 2019 über europäische Unternehmensstatistiken, zur Aufhebung von zehn Rechtsakten im Bereich Unternehmensstatistiken (Amtsblatt der EU Nr. L 327, Seite 1).

Verordnung über die Berichterstattung von Versicherungsunternehmen gegenüber der Bundesanstalt für Finanzdienstleistungsaufsicht (Versicherungsberichterstattungs-Verordnung – BerVersV) vom 19. Juli 2017 (BGBl. I Seite 2858), die durch Artikel 7 des Gesetzes vom 17. August 2017 (BGBl. I Seite 3214) geändert worden ist.

Verordnung über die Rechnungslegung der Kreditinstitute, Finanzdienstleistungsinstitute und Wertpapierinstitute (Kreditinstituts-Rechnungslegungsverordnung – RechKredV) in der Fassung der Bekanntmachung vom 11. Dezember 1998 (BGBl. I Seite 3658), die zuletzt durch Artikel 25 Absatz 6 des Gesetzes vom 7. August 2021 (BGBl. I Seite 3311) geändert worden ist.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im April 2024
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-24002-4, ISSN 1619-2907