



FORSCHUNGSDATENZENTRUM
der Bundesagentur für Arbeit im Institut für
Arbeitsmarkt- und Berufsforschung

ABSCHLUSSBERICHT

Anonymisierungskonzept für einen verknüpften Betriebsdatensatz aus Daten des Bestandes B 412 (THA/BvS) des Bundesarchivs mit den Daten des IAB

Jan Mackeben (IAB), Sandra Dummert (IAB), Heiko Stüber (IAB), Dana Müller (IAB)

Anonymisierungskonzept für einen verknüpften Betriebsdatensatz aus Daten des Bestandes B 412 (THA/BvS) des Bundesarchivs mit den Daten des IAB

Jan Mackeben (IAB), Sandra Dummert (IAB), Heiko Stüber (IAB), Dana Müller (IAB)

Dieser Abschlussbericht beschreibt das Anonymisierungskonzept für einen verknüpften Betriebsdatensatz aus Daten des Bestandes B 412 (THA/BvS) des Bundesarchivs mit administrativen Daten des IAB. Das Projekt wurde im Rahmen der Ausschreibung „Fördermittel Forschungsdatenmanagement“ des Konsortiums für die Sozial-, Verhaltens-, Bildungs- und Wirtschaftswissenschaften (KonsortSWD) in der Nationalen Forschungsdateninfrastruktur (NFDI) gefördert.

Inhaltsverzeichnis

1	Zusammenfassung	3
2	Allgemeine Angaben	4
3	Einleitung	4
4	Die Daten der THA/BvS	7
5	Anonymisierungskonzept	9
5.1	Auswahlschritt 1	9
5.2	Auswahlschritt 2	9
6	Verknüpfung der THA/BvS-Daten mit den Daten des IAB	14
7	Verfügbarkeit der Daten	15
8	Literatur	16

Tabellenverzeichnis

Tabelle 1: Verfügbare Datensätze (Auswahl)	8
Tabelle 2: Wichtigste IDs und deren Häufigkeit	9
Tabelle 3: Ausgewählte Variablen im 2. Schritt	11

1 Zusammenfassung

Ziel des Forschungsprojektes „Anonymisierungskonzept für einen verknüpften Betriebsdatensatz aus Daten des Bestandes B 412, Treuhandanstalt / Bundesanstalt für vereinigungsbedingte Sonderaufgaben (THA/BvS), des Bundesarchivs (BArch) mit den Daten des IAB“ war die Erstellung eines Konzeptes, wie verknüpfte Daten beider Institutionen - BArch und IAB - anonymisiert werden sollten, um diese datenschutzkonform im Sinne des Bundesarchivgesetzes (BArchG) und des Sozialgesetzbuches (SGB) zukünftig der Forschungsgemeinschaft anbieten zu können. Bei den Daten handelt es sich zum einen um Daten der THA/BvS, die beim BArch aufbereitet unter der Signatur B 412/135408 archiviert sind. Diese wurden gem. BArchG dem Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung (FDZ-BA/IAB) für ein Kooperationsprojekt mit dem Institut für Wirtschaftsforschung Halle (IWH) zur Verfügung gestellt und zum anderen um administrative Betriebsdaten des IAB, genauer das Betriebs-Historik-Panel (BHP) ergänzt. Die Daten beider Institutionen verfügen über ähnliche Informationen und sind damit geeignet einen historischen Forschungsdatensatz aufzubauen. Zur Erstellung des Anonymisierungskonzeptes wurden die weitreichenden Daten der THA/BvS durch das IAB zunächst gesichtet. Im Zuge von zwei Auswahlritten erfolgte anschließend die Auswahl geeigneter vergleichbarer Variablen. Während im ersten Auswahlritt die (hypothetische) Eignung für die Arbeitsmarkt- und Berufsforschung – im Sinne des eingeschränkten zweckgebundenen Datenzugangs nach dem SGB – geprüft wurde, bildeten berechnete (Verteilungs-)Kennziffern die Entscheidungsgrundlage des zweiten Auswahlritttes. Informationen, die eine Identifikation von Betrieben erleichtern (oder gar ermöglichen), wurden in beiden Auswahlritten ausgeschlossen. Um eine Re-Identifikation von Betrieben auszuschließen, sollen einige ausgewählte Variablen der Arbeitsmarkt- und Berufsforschung nur in vergrößerter Form bereitgestellt werden und es bedarf nach dessen Aufbereitung einer Stichprobenziehung des verknüpften Datensatzes. Das Anonymisierungskonzept ist ein entscheidender Baustein für die Einhaltung der Datenschutzanforderungen sowohl gem. BArchG als auch gem. SGB und unterstützt die weiteren Arbeiten, ein verknüpftes Datenprodukt aus den THA/BvS-Daten und den administrativen Betriebsdaten des IAB für die Forschungsgemeinschaft zu ermöglichen.

2 Allgemeine Angaben

Das Projekt wurde von Sandra Dummert, Dana Müller und Heiko Stüber (alle FDZ-BA/IAB) mit Unterstützung des BArch, insbesondere durch Jana Michaelis (Leiterin des Projekts „Aufarbeitung der Überlieferung der Treuhandanstalt/Bundesanstalt für vereinigungsbedingte Sonderaufgaben“) und das Referat B 4 (u. a. zuständig für die Archivierung von Daten aus behördlichen Fachverfahren), im Rahmen der Ausschreibung des KonsortSWD „Fördermittel Forschungsdatenmanagement“ im April 2021 beantragt. Die Bewilligung des Projektes erfolgte im Juni 2021 und umfasste eine halbe wissenschaftliche Stelle von 3 Monaten mit einem maximalen Fördervolumen von 11.529 Euro. Als Mitarbeiter im Projekt war Jan Mackeben angestellt. Der Abschlussbericht wurde von Seiten des BArch geprüft.

3 Einleitung

Im Zuge der sich anbahnenden Währungs-, Wirtschafts- und Sozialunion zwischen der Deutschen Demokratischen Republik (DDR) und der Bundesrepublik Deutschland (BRD) stellte sich – weltweit erstmals – die Frage, wie Unternehmen einer Planwirtschaft in das Wirtschaftssystem (soziale Marktwirtschaft) eines anderen Staates überführt werden können. Zur Bewältigung dieser Aufgabe verabschiedete die letzte Volkskammer der DDR am 17. Juni 1990 das Gesetz zur Privatisierung und Reorganisation des volkseigenen Vermögens. Dieses – heute als Treuhandgesetz bekannte – Gesetz verfolgte drei Hauptabsichten:

- 1) die unternehmerischen Tätigkeiten des Staates durch Privatisierung so rasch und so weit wie möglich zurückzuführen,
- 2) die Wettbewerbsfähigkeit möglichst vieler Unternehmen herzustellen und somit Arbeitsplätze zu sichern und neue zu schaffen und
- 3) Grund und Boden für wirtschaftliche Zwecke bereitzustellen.

Mit der Umsetzung dieser Aufgaben wurde (insbesondere) die am 1. März 1990 gegründete Anstalt zur treuhänderischen Verwaltung des Volkseigentums (kurz Treuhandanstalt, THA) beauftragt, die ihre Aufgaben ab dem 1. Januar 1995 unter dem Namen Bundesanstalt für vereinigungsbedingte Sonderaufgaben (BvS) fortführte.

In der verhältnismäßig kurzen Zeit des Bestehens der Behörde sind in der THA/BvS über 45 km Schriftgut entstanden, von denen das Bundesarchiv etwa 170.000 Akten übernommen hat (Bestand B 412). Es setzte sich innerhalb der THA jedoch schnell die Erkenntnis durch, dass

auch auf die Verwendung von Fachverfahren (datenbankbasierten Fachanwendungen) für die Arbeitsorganisation nicht verzichtet werden konnte: „Zur ordnungsgemäßen und effizienten Wahrnehmung der Aufgaben der Treuhandanstalt ist angesichts der erheblichen abzuwickelnden bzw. zu betreuenden Mengengerüste ein einheitliches, weitgehend DV-gestütztes Berichtswesen zwingend erforderlich.“¹ In verschiedenen Datenbanken sammelte die THA/BvS umfangreiche Daten über die Unternehmen der ehemaligen DDR und ihre Entwicklung nach 1990. Die vom Bundesarchiv übernommenen Datenbanken spiegeln den Stand vom 30. Dezember 2008 wider.

Das Bundesarchiv hat die Aufgabe, Archivgut des Bundes auf Dauer zu sichern, nutzbar zu machen und wissenschaftlich zu verwerten. Es bewahrt neben Unterlagen in Form von Akten, Karten, Filmen und Tonaufzeichnungen, die in analoger und digitaler Form vorliegen und teilweise weit in die Vergangenheit (15. Jahrhundert) zurückreichen, auch genuin digitale Unterlagen aus der jüngsten Vergangenheit, insbesondere E-Akten und Daten aus Fachverfahren. Zugang zum Archivgut für eine Nutzung hat grundsätzlich jede Person unter Wahrung des Schutzes privater oder öffentlicher Belange (BArchG § 10).

Insbesondere die Daten zum Arbeitsmarkt der ehemaligen DDR und die Daten aus den Fachverfahren der THA/BvS zur Abwicklung und Überführung der sozialistischen Betriebe in die Marktwirtschaft nach 1990 sind für die Arbeitsmarkt- und Berufsforschung von besonderer Bedeutung, um die Entwicklungen der Betriebslandschaft und die damit verbundenen (sozialen) Ungleichheiten zu untersuchen. Das Bundesarchiv bewahrt dauerhaft die Daten aus den Fachverfahren der THA/BvS, um sie für Nutzungen bereitzustellen. Ebenso wie die Akten stehen auch die Daten der THA/BvS auf Grundlage des BArchG grundsätzlich jedem zur Einsicht zur Verfügung, wenn dadurch nicht die Rechte Dritter oder das Recht auf Geheimhaltung verletzt werden. Auf einem Großteil der Daten liegt wie auch auf den meisten Akten der THA/BvS aufgrund des jungen Entstehungszeitraums noch eine Schutzfrist (BArchG § 11), die unter Wahrung der oben erwähnten Rechte verkürzt werden kann. Dafür werden ggf. potentielle Betriebs- oder Geschäftsgeheimnisse aus den entsprechenden Datensätzen entfernt. Das FDZ-BA/IAB steht seit längerem mit dem Bundesarchiv im Austausch, da sich die Daten der THA/BvS mit denen des IAB inhaltlich gut ergänzen und die Datenstrukturen Ähnlichkeiten aufweisen, sodass sich die Daten sinnvoll verknüpfen lassen.

Der vorliegende Bericht beschreibt die Entwicklung eines Anonymisierungskonzeptes für einen sich im Aufbau befindenden Datensatz. Dieser Datensatz wird verknüpfte Informationen

¹Treuhand Dokumentation 1990-1994, Bd. 2, S. 555

aus den administrativen Betriebsdaten des IAB und den Daten der THA/BvS des Bundesarchivs enthalten.

Das Bundesarchiv hat nach einer Schutzfristverkürzung und in Zusammenarbeit mit dem IWH die THA/BvS-Daten für die Verknüpfung mit administrativen Daten des IAB zur Verfügung gestellt. Zunächst erfolgt eine Verknüpfung der THA/BvS-Daten mit dem BHP (Ganzer et al. 2020) des IAB. Das BHP ist ein Betriebsdatensatz, der neben Informationen zum Wirtschaftszweig sowie der regionalen Zuordnung des Betriebes auf Kreisebene unter anderem Angaben über die Beschäftigtenstruktur als auch aggregierte Angaben zum Tagesentgelt der Vollzeitbeschäftigten enthält. Das BHP setzt sich aus Querschnittsdatensätzen ab dem Jahr 1975 für Westdeutschland und ab 1992 für das vereinigte Deutschland zusammen. Die einzelnen Querschnittsdatensätze können zu einem Panel verbunden werden. Jeder Querschnitt umfasst alle Betriebe, die zum Stichtag 30. Juni in der IAB Beschäftigten-Historik (BeH) erfasst sind. Aufgrund der Einführung des Meldeverfahrens in den neuen Bundesländern kann bei BeH-Sätzen jedoch erst ab 1993 von einer hinreichenden Vollständigkeit der Ost-Meldungen ausgegangen werden. Daher zielt das FDZ-BA/IAB darauf ab, mit Hilfe der Daten der THA/BvS:

- a) ehemalige DDR-Unternehmen bzw. deren Betriebe, die durch die THA/BvS privatisiert wurden, in den administrativen Daten zu identifizieren und
- b) Lücken in den administrativen Daten des IAB zu schließen.

Ziel ist es, die verknüpften Daten der internationalen Forschungsgemeinschaft über das FDZ-BA/IAB für eine Sekundärnutzung mittelfristig zur Verfügung zu stellen. Da ein solches Datenprodukt bisher noch nicht existiert, aus der Forschung allerdings immer wieder Nachfragen zu administrativen Daten über Unternehmen/Betriebe der ehemaligen DDR an das IAB herangetragen werden, erwarten wir eine umfangreiche Nachfrage. Dieser verknüpfte Datensatz würde die einmalige Möglichkeit bieten, nähere Informationen über den Verbleib und die Fortentwicklung ehemaliger DDR-Unternehmen zu erhalten. Forschende könnten anhand dieser Daten identifizieren, welche Faktoren sich auf das Fortbestehen von Unternehmen positiv ausgewirkt haben und welche Unternehmensmerkmale zu frühzeitigen Liquidierungen geführt haben. Von besonderem Forschungsinteresse können auch Vergleiche zwischen der DDR und der BRD sein.

Der Bericht gliedert sich wie folgt: Kapitel 4 stellt die THA/BvS-Daten vor und Kapitel 5 beschreibt die Entwicklung des Anonymisierungskonzeptes. Kapitel 6 geht auf die Verknüpfung der THA/BvS-Daten mit den administrativen Daten des IAB ein und Kapitel 7 auf die geplante Bereitstellung der Daten über das FDZ-BA/IAB.

4 Die Daten der THA/BvS

Die THA/BvS-Daten, die im Bundesarchiv verwahrt werden, wurden nach der Prüfung auf rechtliche Belange und erfolgter Schutzfristverkürzung zunächst an das IWH übermittelt, wobei Daten mit potentiellen Betriebs- und Geschäftsgeheimnissen entfernt wurden. Im IWH wurden die Daten teilweise aufbereitet und im Zuge des Projektes „Entwicklung der ostdeutschen Betriebe nach der Wiedervereinigung“ an das IAB weitergeleitet. Besonders zu erwähnen ist hierbei, dass die THA/BvS ihre Daten in Fachverfahren und Datenstrukturen erfasst und verwaltet, die stetig weiterentwickelt wurden. Dies hat einerseits zu einer sehr umfangreichen, teilweise aber auch heterogenen Informationsmenge geführt. Insgesamt übergab das IWH 86 eigens aufbereitete Datensätze aus den THA/BvS-Daten mit insgesamt 3.152 Variablen an das IAB. Diese gelieferten Datensätze decken ein sehr breites Themenspektrum ab: von Basis Kennziffern (u. a. Kapital, Umsatz, Beschäftigtenanzahlen) und Besitzinformationen (u. a. Eigner, Anteile des Eigners) über Daten zur Betriebsauflösung (u. a. Schwierigkeitsgrad, Status der Abwicklung) bis hin zu Angaben zu Sozialplänen (u. a. Antragsdatum, Antragsvolumen) und Vertragsverhandlungen (Verkauf eines Betriebes, eines Grundstücks oder Inventars etc.). Eine Auswahl gelieferter Datensätze (inklusive enthaltener Informationen) ist in Tabelle 1 ersichtlich.

Der Datensatz „ident_data_tha_id“ enthält beispielsweise Informationen zur relevanten Identifikationsnummer, den Namen, den Produktionsgegenstand, die Registernummer, den Registerort, das Registerdatum, das Gründungsdatum, das Beendigungsdatum, den Liquidationsstatus und das Bundesland. Der Füllgrad dieser Merkmale ist jeweils als gut anzusehen. Der Datensatz umfasst insgesamt 15.431 Unternehmen.

Tabelle 1: Verfügbare Datensätze (Auswahl)

#	Datensatz	Informationen (Auswahl)
1	ident_data_tha_id	IDs, Name und Sitz der Firma
2	basic_data_tha_id	Kapital, Umsatz, Beschäftigung (gesamt und pro Bereich)
3	ownership_data	Eigner, Anteile des Eigners
4	tha_personnel_data	Name, Beruf, Anschrift der Ansprechpersonen
5	tha_id_products	Produktgruppe, Anteil am Firmenumsatz
6	investment_commitments	Fälligkeit und Paragraph der Investitionszusagen
7	investment_audits	Istbetrag und Stichtag Investitionszusagen
8	contract_management_profits	Übernahme v. Verlusten durch THA/BvS, Darlehenslaufzeit
9	ident_data_partial_contracts	Abgang Vermögenskreis, Abgang Sondervermögen
10	social_plans	Antragsdatum, Antragsvolumen
11	labor_commitments	Anzahl und Fälligkeit vereinbarte Arbeitsplätze
12	labor_audits	Arbeitsplätze gemeldet, Ist-Anzahl Arbeitsplätze
13	ident_data_contracts	Einstufung des Vertrages, Angaben zum Notar
14	ident_data_partial_contracts	Fixer Kaufpreis, Eintragung im Grundbuch
15	liquidations	Schwierigkeitsgrad, Status der Abwicklung
16	board_information	Zugesicherte Arbeitsplätze
17	contract_partners	Angaben zu den Vertragspartnern
18	liabilities	Informationen zu Verbindlichkeiten
19	old_liabilities	Information zu Altkrediten
20	resales	Weiterverkauf, Objekt, Käufer
21	sap_guarantees	Informationen zu Bürgschaften
22	meta_data_concentrated_action	Überschuldung, Illiquidität
23	tha_id_succession	Firmenname vor Spaltung/Fusion
24	gdr_plants	Kreis- und Gemeindeschlüssel der Arbeitsstätte

Zwischen den 86 gelieferten Datensätzen bestehen (teilweise) erhebliche Unterschiede hinsichtlich der Anzahl der enthaltenen Unternehmen, der Befüllung von Zellen (Anzahl der fehlenden Werte) oder des Zeitbezuges der Daten und des Formates (Quer- vs. Längsschnitt). Zudem gibt es unterschiedliche Identifikationsnummern (ID) in den Datensätzen (siehe Tabelle 2). Momentan liegt uns nur ein Schlüssel für die Zuordnung der sysnr_id zur tha_id vor.

Tabelle 2: Wichtigste IDs und deren Häufigkeit

#	IDs	Vorhanden in Datensätzen	Inhalt
1	tha_id	32	Firmenschlüssel
2	tha_id_old	6	Alter Firmenschlüssel
3	sysnr_id	47	Schlüssel für Vertragsverhandlung (abgebrochen & erfolgreich)
4	id_partial_contract	36	Schlüssel für (Teil-) Verträge
5	sbr_id	10	Registernummer des Betriebes im DDR-Register

5 Anonymisierungskonzept

Die folgenden Unterkapitel beschreiben das Vorgehen bei der Erstellung des Anonymisierungskonzeptes für die THA/BvS-Daten. Während sich der erste Auswahlschritt mit der Informationssammlung befasste, wurden die im ersten Schritt ausgewählten Variablen im zweiten Auswahlschritt genauer analysiert.

5.1 Auswahlschritt 1

Im ersten Auswahlschritt wurde zunächst eine Excel-Datei mit Informationen zu allen gelieferten Variablen (3.152) erstellt. Anschließend wurden die Variablen von den Autorinnen und Autoren jeweils einer von vier Aufnahmekategorien (1 = „Ja“, 2 = „Eher ja“, 3 = „Unklar“, 4 = „Nein“) zugeordnet. Die Entscheidungsgrundlage für die Aufnahme bildete die (hypothetische) Eignung für die Arbeitsmarkt- und Berufsforschung, denn der Zugang zu den Daten des FDZ-BA/IAB ist rechtlich nach §282 SGB III Abs. 7 und §75 SGB X zweckgebunden für die Arbeitsmarkt- und Berufsforschung und die Forschung im Sozialleistungsbereich. Im ersten Auswahlschritt wurden zudem alle vertraulichen Informationen wie Namen oder Datumsangaben, die ein mögliches Re-Identifikationsrisiko auf Unternehmen/Betriebe oder Eigentümer abbilden, für eine Aufnahme in den Datensatz ausgeschlossen.

In einem nächsten Schritt wurde für alle Variablen, die von mindestens einem Teammitglied nicht der Kategorie „Nein“ zugeordnet wurden, eine Aufnahme in den zweiten Auswahlschritt diskutiert.

5.2 Auswahlschritt 2

Für alle aus dem ersten Auswahlschritt hervorgegangenen Variablen wurden im zweiten Auswahlschritt verschiedene Kennziffern berechnet. Beispiele sind die Anzahl der

Beobachtungen (insgesamt), die Anzahl der Unternehmen, zu denen bei der genannten Variablen mindestens eine Beobachtung vorliegt und Verteilungskennziffern, wie der Mittelwert und verschiedene Perzentile. Anhand dieser Kennziffern wurde in einer weiteren Auswahlrunde die Aufnahme in den finalen Datensatz diskutiert. Im Zuge dieser zweiten Auswahlrunde wurden die Variablen wieder einer von vier Aufnahmegruppen zugeteilt (1 = „Ja“, 2 = „Eher ja“, 3 = „Unklar“, 4 = „Nein“). Die drei wichtigsten Kriterien in der zweiten Auswahlrunde bildeten die Eignung für die Arbeitsmarkt- und Berufsforschung, die Anzahl der Unternehmen zu der mindestens eine Beobachtung vorliegt und die Anzahl der Beobachtungen insgesamt. Tabelle 3 enthält alle Variablen, die den Status 1 = „Ja“ in der zweiten Auswahlrunde erhielten und damit in den finalen Datensatz aufgenommen werden sollen. Während die Spalten 2 und 3 den Namen und das Label der Variablen enthalten, ist in Spalte 4 der Variableninhalt für Variablen ohne Label aufgeführt. Bestimmte Merkmale, die eine Identifikation von Unternehmen erleichtern, sollen nur in vergrößerter Form in den finalen Datensatz aufgenommen werden. Die Spalten 5 und 6 informieren darüber, ob die Variable in vergrößerter Form herausgegeben werden soll. Beispiele für Variablen, die in vergrößerter Form herausgegeben werden sollen, sind der Umsatz und das Kapital. Einige Variablen – wie die Branche und Beschäftigtenzahlen – sollen aus verschiedenen Datensätzen aufgenommen werden, um fehlende Werte in einer Quelle (Datensatz) durch vorhandene Werte in der anderen Quelle (Datensatz) auszugleichen.

Tabelle 3: Ausgewählte Variablen im 2. Schritt

#	Name	Label	Inhalt	Anonymisierung	Codierung
1	tha_id	THA-Nr. als Identifikationsmerkmal	/	Für das Datenprodukt wird eine neue Identifikationsnummer erstellt. Eine Schlüsseltable wird am FDZ-BA/IAB vorgehalten.	Es wird eine neue Zufallsnummer generiert.
2	Gruend_dat	/	Gründungsdatum des Unternehmens	Das Merkmal wird als sensibel eingestuft. Im BHP ist eine vergleichbare Variable vorhanden, die das tagesgenaue erste Auftreten der Nummer des Betriebes enthält, was aber nicht zwangsläufig dem Gründungsdatum entsprechen muss. Da es sich hier um das tatsächliche Gründungsdatum handelt, soll dieses nur auf gesonderten Antrag hin verfügbar gemacht werden. Ansonsten soll nur das Gründungsjahr zur Verfügung gestellt werden.	Nur das Jahr sollte im Datensatz ausgewiesen werden. Das Datum kann bei begründeter Erforderlichkeit zur Verfügung gestellt werden.
3	Taetenddat	/	Beendigungsdatum des Unternehmens	Das Merkmal wird als sensibel eingestuft. Im BHP ist eine vergleichbare Variable vorhanden, die das tagesgenaue letzte Auftreten der Nummer des Betriebes enthält, was aber nicht zwangsläufig dem Datum der Betriebsschließung entsprechen muss. Da es sich hier um das tatsächliche Beendigungsdatum handelt, soll dieses nur auf gesonderten Antrag hin verfügbar gemacht werden. Ansonsten soll nur das Beendigungsjahr zur Verfügung gestellt werden.	Nur das Jahr sollte im Datensatz ausgewiesen werden. Das Datum kann bei begründeter Erforderlichkeit zur Verfügung gestellt werden.
4	Statusn	/	Status der Abwicklung	Die Variable hat viele Ausprägungen und sollte vergrößert werden. Vier Gruppen werden vorgeschlagen.	1 Bestehendes Unternehmen: Im Besitz der THA/BvS 2 Bestehendes Unternehmen: Nicht im Besitz der THA/BvS 3 Liquidiert, fusioniert, aufgelöst

			(Liquidation/Verkauf/Privatisierung)		4 unklare Eignerverhältnisse
5	Stamm_kapital	Nennkapital in TOM (1989: TM)	/	Einstufung als sensibles Merkmal	Das Merkmal sollte im Datensatz vergrößert ausgewiesen werden. Vorschlag: Gruppierung anhand der 10er Perzentile. Bei begründeter Erforderlichkeit kann die originäre Information zur Verfügung gestellt werden.
6	Umsatz	Jahresumsatz in TOM (1989: TM)	/	Einstufung als sensibles Merkmal	Das Merkmal sollte im Datensatz vergrößert ausgewiesen werden. Vorschlag: Gruppierung anhand der 10er Perzentile. Bei begründeter Erforderlichkeit kann die originäre Information zur Verfügung gestellt werden.
7	Besch_gesamt	Beschäftigte gesamt	/	Die Variable ist im BHP nicht vergrößert	Identisches Vorgehen wie im BHP
8	Besch_finan_z_rewe	Beschäftigte im Finanz-/ Rechnungswesen	/	Die Variable ist im BHP nicht vergrößert	Identisches Vorgehen wie im BHP
9	Besch_einkauf	Beschäftigte im Einkauf	/	Die Variable ist im BHP nicht vergrößert	Identisches Vorgehen wie im BHP
10	Besch_pers_sozialw	Beschäftigte im Personal-/ Sozialwesen	/	Die Variable ist im BHP nicht vergrößert	Identisches Vorgehen wie im BHP
11	Besch_produktion	Beschäftigte in der Produktion	/	Die Variable ist im BHP nicht vergrößert	Identisches Vorgehen wie im BHP
12	Besch_f_e	Beschäftigte in Forschung und Entwicklung	/	Die Variable ist im BHP nicht vergrößert	Identisches Vorgehen wie im BHP
13	Besch_vertrieb	Beschäftigte im Vertrieb	/	Die Variable ist im BHP nicht vergrößert	Identisches Vorgehen wie im BHP

14	Besch_verwal_sonst	Beschäftigte in Verwaltung und sonstige	/	Die Variable ist im BHP nicht vergrößert	Identisches Vorgehen wie im BHP
15	Branche	Branchen-Schlüssel	/	Branchen auf 4 -Steller Ebene	Aggregation auf 2-Steller-Ebene bei zu geringen Fallzahlen im Endprodukt notwendig.
16	Bundesland	Abkürzung für Bundesland	/	Zusammenfassung der westdeutschen Bundesländer zu einer Kategorie	Identisches Vorgehen wie im BHP, außer für die westdeutschen Bundesländer, die zusammengefasst werden sollten

6 Verknüpfung der THA/BvS-Daten mit den Daten des IAB

Die Verknüpfung der THA/BvS-Daten mit den administrativen Betriebsdaten des IAB erfolgt mit Hilfe von Record-Linkage-Methoden. Als betriebsidentifizierende Merkmale wurden der Name des Unternehmens, die Rechtsform und der Ort herangezogen. Um eine möglichst hohe Übereinstimmung zwischen beiden Datenquellen zu erreichen, wurden die betriebsidentifizierenden Merkmale zunächst Bereinigungsverfahren unterzogen. Konkret wurden Umlaute und Sonderzeichen durch ihre jeweiligen ASCII-konformen Äquivalente ersetzt, alle Buchstaben in Großbuchstaben umgewandelt, überflüssige Leerzeichen und Interpunktionszeichen entfernt sowie Abkürzungen ausgeschrieben und Schreibweisen vereinheitlicht.

Nach Bereinigung und Standardisierung der Namens- und Ortskomponenten wurde zunächst ein deterministischer Abgleich der Felder zwischen den Einträgen der beiden Datenquellen durchgeführt. Im ersten deterministischen Linkage-Schritt war die exakte Übereinstimmung in allen zur Verfügung stehenden Identifikatoren ausschlaggebend, also Betriebsname, Rechtsform und Ort. In einem nächsten Schritt wurde bei nicht verknüpften Betrieben eine fehlende Rechtsform toleriert, während die jeweils übrigen Felder weiterhin exakt übereinstimmen mussten.

Für Unternehmen, die weiterhin nicht in den administrativen Daten gefunden werden konnten, wurde anschließend ein fehlertolerantes Matching-Verfahren verwendet. Dabei wurde für jedes abzugleichende Variablenpaar aus beiden Datenquellen ein Ähnlichkeitsmaß berechnet und diese in einem zusammenfassenden Qualitätsindex aufsummiert. Die Berechnungen wurden mit der Software Merge ToolBox (MTB, Version 0.751, siehe Schnell et al., 2005; Bachteler, 2012) durchgeführt. Datenpaare, deren Ähnlichkeitsmaß unter einem durch vorläufige Sichtung bestimmten Schwellenwert lag, wurden gelöscht. Zuordnungen über diesem Schwellenwert wurden anschließend einem Sichtvergleich unterzogen.

Anschließend konnten den durch das Record-Linkage erfolgreich verknüpften THA/BvS-Unternehmen die Betriebsinformationen aus dem BHP hinzugespielt werden. Die Verknüpfung der Daten ist noch nicht abgeschlossen. Zum Zeitpunkt der Erstellung des Abschlussberichtes konnten 5.619 THA-IDs (tha_id) mit 6.832 Betriebsnummern (betnr) verlinkt werden.

7 Verfügbarkeit der Daten

Für die Forschungsgemeinschaft stellt sich sicher die Frage, wann der verknüpfte Datensatz der THA/BvS-Daten mit den administrativen Daten des IAB über das FDZ-BA/IAB veröffentlicht wird und wie dieser beantragt werden kann. Ein genaues Veröffentlichungsdatum steht aktuell noch nicht fest. Ein Zugang zu den im Datensatz verwendeten THA/BvS-Daten aus dem Bestand B 412 des BArch kann dort beantragt werden. Es gelten die üblichen Zugangsvoraussetzungen für Archivgut gem. BArchG.

Als Zugangswege für den Datensatz des IAB sind der Gastaufenthalt und die kontrollierte Datenfernverarbeitung im FDZ-BA/IAB vorgesehen. Aufgrund noch anstehender umfangreicher Aufbereitungsschritte aus Eigenmitteln am FDZ-BA/IAB und der konzeptionellen Erarbeitung der Stichprobenziehung wird eine Veröffentlichung des Datensatzes für das Jahr 2023 angestrebt.

8 Literatur

- Bachteler, Tobias (2012): Merge ToolBox – MTB: Record Linkage Software, Version 0.742, Getting Started. Universität Duisburg-Essen. URL: http://www.uni-due.de/~hq0215/documents/mtb_gettingstarted.pdf (Zugriff: 21.03.2022).
- Bundeszentrale für politische Bildung (2020): Die Geburtsstunde der Treuhand. URL: <https://www.bpb.de/kurz-knapp/hintergrund-aktuell/201919/die-geburtsstunde-der-treu-hand/> (Zugriff: 18.03.2022)
- Die Bundesregierung (2022): Volkskammer verabschiedet Treuhandgesetz. URL: <https://www.bundesregierung.de/breg-de/aktuelles/volkskammer-verabschiedet-treuhand-gesetz-406354> (Zugriff: 18.03.2022)
- Ganzer, Andreas; Schmidlein, Lisa; Stegmaier, Jens; Wolter, Stefanie (2020): Betriebs-Historik-Panel 1975-2019. FDZ-Datenreport, 16/2020 (de), Nürnberg.
- Schnell, Rainer; Bachteler, Tobias; Reiher, Jörg (2005): MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung. In: ZA-Information/Zentralarchiv für Empirische Sozialforschung, Nr. 56, S. 93-103.