Bertrand Lisbach

Linguistisches Identity Matching

Bertrand Lisbach

Linguistisches Identity Matching

Paradigmenwechsel in der Suche
und im Abgleich von Personendaten

Mit 9 Abbildungen und 20 Tabellen

PRAXIS

VIEWEG+
TEUBNER

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über
<http://dnb.d-nb.de> abrufbar.

Höchste inhaltliche und technische Qualität unserer Produkte ist unser Ziel. Bei der Produktion und Auslieferung unserer Bücher wollen wir die Umwelt schonen: Dieses Buch ist auf säurefreiem und chlorfrei gebleichtem Papier gedruckt. Die Einschweißfolie besteht aus Polyäthylen und damit aus organischen Grundstoffen, die weder bei der Herstellung noch bei der Verbrennung Schadstoffe freisetzen.

1. Auflage 2011

Lektorat: Christel Roß | Walburga Himmel

Vieweg+Teubner Verlag ist eine Marke von Springer Fachmedien.
Springer Fachmedien ist Teil der Fachverlagsgruppe Springer Science+Business Media.
www.viewegteubner.de

Umschlaggestaltung: KünkelLopka Medienentwicklung, Heidelberg
Umschlagmotiv: artbeat graphic design, Bern (Schweiz)
Druck und buchbinderische Verarbeitung: STRAUSS GmbH, Mörlenbach
Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.
Printed in Germany

ISBN 978-3-8348-1371-8

Geleitwort

Von Dr. Klaus Frick

Seit über 15 Jahren unterstütze ich weltweit agierende Firmen darin, ihre global verteilten Datenbestände effektiv zu nutzen. Die Herausforderungen – organisatorische, technische, fachliche – sind über diesen Zeitraum grundsätzlich dieselben geblieben und übrigens auch die konzeptionellen Lösungsansätze. Die Datenvolumina nehmen kontinuierlich zu und noch mehr tut dies die Geschwindigkeit, mit der diese prozessiert werden können. Aber grundlegend hat sich meine Arbeit in den letzten 15 Jahren nicht geändert und ich denke auch nicht, dass sich daran in Kürze etwas ändern wird. Allerdings wird in Zukunft die Notwendigkeit des aktiven Datenmanagements der global verteilten Datenbestände u.a. aus wirtschaftlichen und regulatorischen Gründen weiter an Bedeutung gewinnen.

Es gehört sicher zu den anspruchsvollsten Aufgaben, heterogene und verteilte Datenbestände in einem global agierenden Unternehmen zu konsolidieren und einer unternehmensweiten Nutzung zuzuführen. Dabei erweisen sich in der Regel zwei Themenbereiche unter fachlichen und organisatorischen Gesichtspunkten als besonders schwierig: ein für das Unternehmen einheitlich gültiger Produktekatalog sowie eine zentralisierte Kundendatenbank, bzw. zumindest die Darstellung einer global gültigen und standardisierten Sicht auf die jeweilig zulässigen lokalen Ausprägungen. Während für die angestrebte Standardisierung des Produktekatalogs sich zwischenzeitlich für die meisten Industrien de-facto-Standards etabliert haben, ist dies in Bezug auf die Kunden- und Personendaten nicht der Fall.

Damit werden Personendaten nun häufig zum Dreh- und Angelpunkt bei meiner Arbeit. Grundsätzliche Anforderung an die Informationssysteme ist, dass diese erkennen, welche Datensätze in den global verteilten Datenbanken dieselben und welche unterschiedliche Personen repräsentieren. Die Rede ist also vom Dublettenerkennen, einer klassischen Aufgabenstellung des Identity Matching. Automatisiertes Dublettenerkennen ist deshalb so schwierig, weil Daten in unterschiedlichen Ländern unterschiedlich eingegeben werden. Insbesondere die Verarbeitung von Namensbestandteilen erweist sich oft als verhängnisvoll: Der Name derselben Person kann aus unterschiedlich vielen Bestandteilen bestehen, die teilweise zusammen oder getrennt geschrieben sind. Die Reihenfolge der Bestandteile kann variieren und auch die Namensdatenstruktur. Mit der Globalisierung der Unternehmen kommt es zwischenzeitlich zunehmend häufiger vor, dass Namen in

unterschiedlichen Alphabeten repräsentiert sind und selbst bei der Darstellung im selben Alphabet auf unterschiedliche Art transkribiert werden. Von Spitznamen, Abkürzungen und Schreibfehlern im Namen ganz zu schweigen.

Inzwischen habe ich gelernt, dass das zuverlässige und präzise Matchen von Namensdaten kein triviales Unterfangen ist und sich bis heute auch noch keine allgemein gültigen Standards etabliert haben. Auf der anderen Seite ist es erstaunlich, wie relativ leicht es jedem Laien fällt, anhand von Beispielnamen korrekt zu bestimmen, ob zwei Namen dieselbe Person referenzieren oder nicht und wie schwer es gleichzeitig den allermeisten Informatiksystemen fällt, diese Aufgabe ebenso zuverlässig zu lösen.

Die Sache ist leicht zu erklären: Wie Bertrand Lisbach in diesem Buch ausführt, war man bisher im falschen Paradigma befangen: Man hat mit technischen Mitteln versucht, das eigentlich linguistische Thema des Identity Matchings in den Griff zu bekommen. Dementsprechend unbefriedigend waren die damit erzielten Ergebnisse.

Aber das ist schon nahezu Vergangenheit. Der Einsatz neu entwickelter Informatiksysteme, die entsprechend dem Paradigma des linguistischen Identity Matching entwickelt wurden, zeigt eine signifikant bessere Trefferquote unter gleichzeitiger Reduktion der unerwünschten Fehltreffer. Nicht nur bei der Dublettenerkennung auch im Compliance-Umfeld von Banken, im Customer Relationenship Management oder bei der Personensuche in Melderegistern oder Telefonverzeichnissen erfahren linguistisch basierte Lösungsvarianten einen zunehmend verbreiteten Einsatz.

Dieses Buch ist das erste, welches sich umfassend mit dem vielseitigen Thema des linguistischen Identity Matching befasst. Bertrand Lisbach hat ein hochinteressantes Buch zu einem Thema mit rapide wachsender Bedeutung verfasst. Für Entscheidungsträger und Experten, aber auch als zeitgemäße Referenz für engagierte Anwender in diesem sehr dynamischen Themenbereich geschrieben, liefert "Linguistisches Identity Matching" ein fundiertes Nachschlagewerk und gleichsam Entscheidungshilfe für den Einsatz von Identity Matching-Methoden für das eigene Unternehmen.

Ich wünsche Ihnen eine spannende Lektüre mit viel Anregung für die Praxis.

Dr. Klaus Frick

Global Datawarehouses Expert

Inhaltsverzeichnis

Einleitung: Paradigmenwechsel im Identity Matching

Ein Elementarprozess in Wirtschaft und Gesellschaft

Die Suche nach Personendaten ist uns allen gut vertraut. Wenn wir uns im Internet Kontaktdaten von Freunden, Arbeitskollegen oder Geschäftspartnern besorgen, starten wir eine *Personensuche*. Dasselbe tun wir, wenn wir aus privatem oder wissenschaftlichem Interesse über eine öffentliche Person recherchieren wollen. Natürlich sind wir selbst auch – und vielleicht weit häufiger als uns lieb ist – Gegenstand einer Personensuche.

Sobald wir uns an einen Bankschalter begeben, um ein Konto zu eröffnen, wird anhand unserer Ausweispapiere festgestellt, ob wir uns auf Listen von bekannten Geldwäschern, sanktionierten Parteien, Terroristen oder politisch exponierten Personen befinden. Geben wir als Neukunden eine Bestellung beim Versandhandel auf oder wollen wir eine neue Wohnung anmieten, ist es heute üblich, dass die Gegenpartei zur Prüfung unserer Bonität eine Personensuche über uns bei einer sogenannten Auskunftei absetzt. Auch die Personalabteilung der Firma, für die wir arbeiten, mag eine Personensuche starten oder der Relationship Manager eines Unternehmens, dessen Kunde wir sind oder – aus Sicht des Unternehmens – werden sollen. Es wird auch jeder zum Zielobjekt einer Personensuche, der auf den Radar von Sozialämtern, Strafverfolgungsbehörden oder gar Geheimdiensten geraten ist.

Solchen Personensuchen strukturell sehr ähnlich ist der *Personendatenabgleich*. Bei der Personensuche werden Personendaten in der Suchanfrage mit jenen des Datenpools verglichen, in welchem gesucht wird. Beim Personendatenabgleich wird aufgrund verschiedener Merkmale, z.B. Name und Geburtsdatum, überprüft, ob zwei im Vergleich stehende Datenprofile in einer Datenbank dieselbe oder zwei verschiedene Personen repräsentieren. In den meisten größeren Organisationen werden Personendaten permanent abgeglichen, z.B. zur Sicherstellung der Datenqualität oder zur Gewährleistung einer ganzheitlichen Kundensicht. Heutzutage beruhen etliche und oft sehr zentrale Geschäftsprozesse auf einem zuverlässigen und präzisen Abgleich von Personendaten. Dieser Abgleich hat inzwischen eine so hohe strategische Bedeutung, dass sich zahlreiche Anbieter von Software und Beratungsdienstleistungen ausschließlich darauf spezialisiert haben. Projekte, welche sich *Master Data Management*, *Customer Data Integration* oder *Daten-Deduplizierung* auf die Fahnen geschrieben haben, sind ohne solche Produkte und Leistungen zum Scheitern verurteilt.

Personensuche und Personendatenabgleich sind also zwei enorm weit verbreitete Funktionen, die das heutige gesellschaftliche und wirtschaftliche Leben prägen.

B. Lisbach, *Linguistisches Identity Matching*, DOI 10.1007/978-3-8348-9791-6_1,

Und beide basieren auf demselben Elementarprozess, dem *Identity Matching*. *Identity Matching* vergleicht Personendaten und ermittelt den Grad der Übereinstimmung, üblicherweise mittels eines *Matchscores*. Je höher der *Matchscore* ist, desto wahrscheinlicher handelt es sich bei den verglichenen Personendaten um dieselbe Person – zumindest in der Theorie.

Die linguistische Herausforderung

In der Praxis hat sich Identity Matching als wesentlich anspruchsvoller herausgestellt, als ursprünglich angenommen. Dafür gibt es zahlreiche Gründe, die in diesem Buch ausführlich behandelt werden. Auf den Hauptgrund sei aber schon an dieser Stelle eingegangen: Die Crux beim Identity Matching liegt im (Personen-) Namen, der nicht nur das mit Abstand wichtigste, sondern auch das mit Abstand herausforderndste Personenmerkmal darstellt.

Warum herausfordernd? Weil der Name ein und derselben Person in einer Suchabfrage, in einem Datensatz, in einem Pass oder in einem Zeitungsartikel, oftmals ganz unterschiedlich repräsentiert werden kann. Dies ist keine neue Erkenntnis. Es liegen für einige der Variationsquellen von Personennamen sogar schon seit geraumer Zeit überzeugende Matchmethoden vor. Diese Matchmethoden können z.B. mit Variationen im Gebrauch von Initialen, in der Reihenfolge der einzelnen Namensbestandteile oder in der Datenstruktur umgehen. Die Rede ist hier von einfachen, nicht-linguistischen Verfahren und Algorithmen für einfache Variationsquellen.

Hier sind drei Beispiele für nicht-linguistische Variationsquellen von Namen:

- Gebrauch von Initialen:
  John Fitzgerald Kennedy im Gegensatz zu *John F. Kennedy*
- Reihenfolge:
  John Fitzgerald Kennedy im Gegensatz zu *Kennedy John Fitzgerald*
- Datenstruktur:
  Vorname *John Fitzgerald* im Gegensatz zu Vorname *John* und Zwischenname *Fitzgerald*.

Es sind allerdings nicht diese eher trivialen Variationsquellen, die das *Name Matching* als Teil des Identity Matching so schwierig gestalten. Die wahre Herausforderung stellen *linguistische Namensvariationen* dar. *Linguistische Namensvariationen* kommen durch linguistische Phänomene zustande, die oftmals komplex sind und in verschiedenen Sprachräumen ganz unterschiedliche Ausprägungen haben können. *Linguistische Namensvariationen* zuverlässig und präzise zu matchen bedarf innovativer linguistischer Methoden.

Ein einfaches Beispiel soll dies veranschaulichen. Der erste Präsident Russlands ist in der englischsprachigen Presse vor allem als *Boris Nikolayevich Yeltsin*, in der französischsprachigen als *Boris Nikolaïevitch Eltsine*, in der portugiesischsprachigen als *Boris Nicoláievitch Ieltsin* und in der deutschsprachigen als *Boris Nikolajewitsch Jelzin* bekannt. Eine einheitliche Schreibweise in lateinischer Schrift existiert also nicht. Stattdessen kommen in den einzelnen Ländern (und übrigens auch innerhalb eines Landes) unterschiedliche Transkriptionsstandards zur Anwendung, die auf je unterschiedliche Weise denselben kyrillischen Namen (hier: Борис Николаевич Ельцин) in das lateinische Alphabet überführen.

Sucht z.B. ein Engländer oder Amerikaner mit dem Namen *Yeltsin* in einer Datenbank, in welcher ein Deutscher den Namen *Jelzin* erfasst hat, sollte das Identity-Matching-Tool einen Treffer produzieren, denn die Suchabfrage und das Datenprofil mögen dieselbe Person referenzieren. Um aber *Jelzin* mit *Yeltsin*, *Eltsine* und mit *Ieltsin* zuverlässig und präzise matchen zu können, müssen vorher die verschiedenen Transkriptionsstandards analysiert und als Regeln oder Algorithmen in die Matchmethode implementiert worden sein – eine typisch (computer-) linguistische Herangehensweise. Mit solchen Matchmethoden wird also nicht *exakt* gesucht, sondern *unscharf* oder englisch: *fuzzy*. Das Ziel einer solchen *unscharfen Suche* ist es, dass der Suchende jede Transkriptionsvariante eines Namens verwenden kann und damit jede andere mögliche Transkriptionsvariante des Namens finden kann. Egal also, ob mit *Jelzin*, *Yeltsin*, *Eltsine* oder *Ieltsin* gesucht wird, die Resultatelisten sollten die gleichen sind.

Neben den hier aufgezeigten unterschiedlichen Transkriptionsstandards existieren viele weitere linguistische Quellen von Namensvariationen. Diese bedürfen spezieller (computer-)linguistischer Methoden. Man denke z.B. an das Phänomen der *Homophonie* (gleichlautende Namensvarianten, z.B. *Meier* und *Mayr*), an die Ableitung von Nicknames aus einer Grundform (z.B. *Bill* und *William*) oder an sprachraumübergreifende Verwandtschaftsbeziehungen von Namen (z.B. *Stefan*, *Stéphane*, *Steven*, *Stefano*, *Estêvão* und *Esteban*).

Den nicht-linguistischen, eher trivialen Variationsquellen kann mit relativ einfachen technischen, nicht-linguistischen Mitteln begegnet werden. Für die viel komplexeren linguistischen Variationsquellen sind hingegen spezielle linguistische Methoden unerlässlich. Genau dies ist der Hintergrund für den gegenwärtig zu beobachtenden *Paradigmenwechsel*, der sich im Identity Matching vollzieht: die Abkehr von einfachen, nicht-linguistischen Matchmethoden und die Hinwendung zu einem *linguistischen Identity Matching*. Natürlich kommen auch beim linguistischen Identity Matching verschiedene nicht-linguistische, z.B. algebraische und probabilistische (an der Wahrscheinlichkeitstheorie orientierte) Verfahren und Algorithmen zur Anwendung. In erster Linie basiert linguistisches Identity Matching jedoch auf der Analyse linguistischer Phänomene, für welche Matchmethoden erst noch erfunden werden mussten – ein innovativer Akt, der

sich aber bereits sehr rasch ausbezahlt hat. Überall dort, wo heute Personensuche und Personendatenabgleich auf linguistischem Identity Matching aufsetzen, zeigt sich dies in einer spürbar gesteigerten Qualität. Dies betrifft sowohl die Präzision des Matching als auch dessen Zuverlässigkeit.

Für wen wurde dieses Buch geschrieben?

Dieses Buch eignet sich sowohl für Berufspraktiker, die bei ihrer Arbeit auf ein effektives Identity Matching angewiesen sind, als auch für Studierende und Lehrende diverser informatiknaher und sprachwissenschaftlicher Fächer.

Falls Ihr Interesse an dem Thema Ihrer beruflichen Tätigkeit entstammt, werden Sie in dem Buch viele Fakten, Konzepte, Ideen und Hilfsmittel finden, welche Sie für Ihre Arbeit direkt gebrauchen können. Dies gilt unabhängig davon, ob Sie ein technisch orientierter Software-Entwickler sind, als Power-User einer operative Tätigkeit nachgehen, als Produkt-Manager oder Business-Analyst einen eher konzeptionellen Zugang haben oder als Führungskraft Identity Matching aus einer strategischen Perspektive betrachten. Das Buch hat keinen besonderen Branchen-Fokus. Viele Anwendungsbeispiele entstammen jedoch der Finanzdienstleistungsbranche, welche traditionell dem Thema Identity Matching besonders viel Beachtung entgegenbringt.

Es sind vor allem Angehörige der hier aufgelisteten Berufsgruppen, die in den folgenden Kapiteln relevante Einsichten für ihre Berufspraxis erwarten dürfen:

- Personen mit Führungs-, Fach- oder Umsetzungsverantwortung in den Compliance-Abteilungen von Banken und Versicherungen.
- Beauftragte für Datenqualität, Daten-Deduplizierung und Master Data Management.
- Produkt-Manager und Business-Analysten, die an der Erstellung und Einführung von Software-Lösungen mitwirken, welche Such- und Abgleichsfunktionen aufweisen. Zu diesen Lösungen zählen z.B. Compliance- und CRM-Suites, Enterprise Search Platforms oder Internet Search Solutions.
- Produkt-Manager und Business-Analysten für Datenprodukte (z.B. Adressdaten, Watchlists, News) und Informationsdienste (z.B. Abfragedienste von Bibliotheken, Archiven oder Auskunfteien).
- Investigativ arbeitende Berufsgruppen, z.B. Wirtschaftsprüfer, Journalisten und Detektive sowie Mitarbeiter von Polizei- und Sozialbehörden, in deren Verantwortungsbereich die informatikgestützte Kriminalitäts- und Missbrauchsbekämpfung fällt.

Diesen Berufsgruppen stand bisher nur Fachliteratur zur Verfügung, in welcher das Thema des Identity Matching als rein technisches und nicht als linguistisches Problem dargestellt wurde. Das vorliegende Buch versucht, den in letzter Zeit mit soviel Erfolg entwickelten linguistischen Ansätzen erstmalig den Platz einzuräumen, welcher ihnen gebührt. Dies tut es durch eine praktische Ausrichtung, die sich konsequent durch alle Kapitel zieht. Nach der Lektüre sollen Sie zu Folgendem fähig sein:

- Sie können ermitteln, mit welcher Zuverlässigkeit und Präzision die von Ihnen entwickelte, eingeführte oder benutzte Identity-Matching-Lösung globale Namensdaten matchen kann (*Ist-Analyse*).
- Sie können genaue Anforderungen an eine Identity-Matching-Lösung stellen, die entweder neu zu erstellen, neu zu erwerben oder weiterzuentwickeln ist (*Soll-Analyse*).
- Sie können sachkundig Evaluationen, Benchmark-Studien und Proof of Concepts von Identity-Matching-Lösungen planen und durchführen. Sie sind besser in der Lage, Behauptungen von Software-Anbietern in ihrer Relevanz einzuschätzen und in ihrer Richtigkeit zu überprüfen.

Das Buch eignet sich darüber hinaus als Lehr- und Studienbuch. Es spricht Lehrende und Studierende der Informatik und Wirtschaftsinformatik, der Informationswissenschaften, der Computerlinguistik, der Sprachwissenschaften und der *Onomastik* (Lehre von Eigennamen) an.

Zum Schluss dieser Einleitung noch ein Hinweis darauf, welche Erwartungen das Buch *nicht* erfüllen kann: Es behandelt nicht die Missbrauchsgefahren, die mit Identity Matching verbunden sind. Juristische, ethische und gesellschaftliche Aspekte bleiben also ausgespart. Sie verdienen eine separate Abhandlung.

Dass wir uns hier ausschließlich auf die Aspekte Linguistik, Technologie und Ökonomie des Identity Matching konzentrieren, bestreitet keineswegs die Relevanz einer Diskussion um Missbrauchsgefahren. Diese ist umso wichtiger, als qualitativ hochstehende Identity-Matching-Lösungen dem, der sie beherrscht, ein effizientes Machtmittel zur Hand geben. Und während diese Technologien ihre Legitimität aus Anwendungen wie der Bekämpfung der organisierten Kriminalität oder der Geldwäsche gewinnen, so dürfen doch die damit einhergehenden Gefahren nicht ignoriert werden. Allem Anschein nach stellt eine missbräuchliche und widerrechtliche Nutzung von privater und staatlicher Seite keine Seltenheit dar.

Teil I: Grundlagen des linguistischen Identity Matching

In der Einleitung haben Sie erfahren, was Identity Matching genau bedeutet, worin der Hauptgrund für die Qualitätsprobleme liegt und was Identity Matching heute zu leisten vermag. Sie haben gelesen, dass Identity Matching lange Zeit als eine technische Herausforderung (miss-)verstanden wurde, der folgerichtig, aber bedauerlicherweise mit rein mathematisch-technischen Lösungen zu begegnen versucht wurde. Es wurde auch deutlich, dass Identity Matching zum wesentlichen Teil Name Matching ist und dass, um globale Personennamen mit all ihren Schreibvariationen zuverlässig und präzise matchen zu können, allen voran linguistische Verfahren aufzubieten sind.

Genau dies ist der *Paradigmenwechsel,* der zurzeit zu beobachten ist und der zu einer sprunghaften Qualitätssteigerung im Identity Matching geführt hat. Die Grundideen und Verfahren, welche sich hinter dem neuen Paradigma verbergen, machen den roten Faden aus, der sich durch dieses Buch zieht. Die Kapitel dieses ersten Teils liefern dafür das fachliche und terminologische Grundgerüst.

Das erste Kapitel gibt eine kurze und allgemein verständliche Einführung in die Grundkonzepte des Identity Matching und in seine wichtigste Komponente, dem Name Matching. Dabei bedient es sich einer dem Information Retrieval entnommenen Terminologie. So wird es möglich, die Zielsetzung von Identity Matching auf die einfache Formel zu bringen: Maximale Trefferquote (Recall) und maximale Genauigkeit (Precision) bei der Personensuche und beim Abgleich von Personendaten. Als Reaktion auf die oft zu lesende Behauptung, dass sich die Trefferquote nur auf Kosten der Präzision, und die Präzision nur auf Kosten der Trefferquote erhöhen ließe, wird aufgezeigt, dass linguistische Methoden einen Ausweg aus dem Dilemma bieten.

Das zweite Kapitel steckt die zahlreichen und vielfältigen Anwendungsfelder von Identity Matching ab. Identity Matching ist ein Elementarprozess, der allein in der Finanzdienstleistungsindustrie Dutzende verschiedener Geschäftsprozesse unterstützt – von der Geldwäsche-, Missbrauchs- und Terrorbekämpfung bis hin zum Customer Relationship Management. Sozialämter dämmen mit Identity Matching Betrügereien ein. Polizeibehörden hilft es im Kampf gegen Kriminalität und bei der Strafverfolgung. Investigativ arbeitende Berufsgruppen, wie Journalisten, Detektive und einige Wirtschaftsprüfer, sind ebenso auf ein qualitativ hochwertiges Identity Matching angewiesen wie Verlage, Nachrichtenagenturen und andere Informationsdienstleister. Nicht zuletzt müssen sich auch IT-Abteilungen von Unternehmen auf die Qualität des Identity Matching verlassen können, wenn sie den Verkaufs-, Marketing- und Controlling-Einheiten ihres Unternehmen be-

reinigte und konsolidierte Daten über Kunden, Mitarbeiter oder Partner bereitstellen wollen.

Die Kapitel 3 bis 6 behandeln die wichtigsten linguistischen Variations- und Fehlerquellen von Namen. Denn nur wenn diese in ihrer Verbreitung und Ausprägung bekannt sind, können sinnvollerweise Anforderungen an das Identity Matching formuliert werden. Ohne linguistische oder namenskundliche Kenntnisse vorauszusetzen, werden in diesen Kapiteln Besonderheiten in der Struktur und der Schreibung von Personennamen geschildert. Dabei nehmen wir eine globale Perspektive ein, denn die Struktur von Namen, die Verwendung der einzelnen Namenselemente, die Aussprache, die originale Schreibweise wie auch die Schreibweise, die resultiert, wenn Namen in das lateinische Alphabet überführt werden, variieren stark zwischen den verschiedenen Sprach- und Kulturräumen. Daher werden Sie nicht nur vieles über westliche Namenssysteme und Sprachen erfahren, sondern auch über Namen z.B. des russischen, arabischen und chinesischen Sprachraums.

Neben den linguistischen Quellen von Schreibvariationen in Namensdaten gibt es auch nicht-linguistische. Diese werden unter dem Stichwort "Tippfehler" in Kapitel 7 behandelt. Es wird aufgezeigt, wie durch die Analyse der verschiedenen Ursachen von Tippfehlern, weit präzisere und zuverlässige Matching- und Suchergebnisse möglich werden als mit herkömmlichen Methoden.

Wenn Sie den ersten Teil gelesen haben, sind Sie terminologisch und fachlich beim Thema des linguistischen Identity Matching auf der Höhe. Sie wissen nicht nur, wo Identity Matching Anwendung findet und was dessen Hauptaufgabe ist, nämlich das zuverlässige und präzise Matchen von Namensvariationen. Sie können darüber hinaus Arten von Variationen und Fehlern in der Schreibung von Namen unterscheiden und kennen deren Ursachen. Damit sind die Grundlagen für Teil II geschaffen, in welchem die gängigen Name-Matching-Verfahren unter die Lupe genommen und verglichen werden.

1 Grundkonzepte

1.1 Identity Matching und Name Matching

Identity Matching ist ein Vorgehen zur Beantwortung der Frage, ob zwei unterschiedliche Datenobjekte dieselbe oder verschiedene Personen repräsentieren. Dies geschieht, indem die Übereinstimmung der Merkmale beider Datenobjekte ermittelt wird. Bei den Datenobjekten handelt es sich typischerweise um Suchabfragen oder um Datenprofile (in Form von Datensätzen oder Indexstrukturen). Bei den Personen, die durch die Datenobjekte repräsentiert werden, kann es sich um natürliche Personen oder um juristische Personen handeln. Juristische Personen sind z.B. Stiftungen, Vereine oder Firmen.

Wenn zwei Datenobjekte dieselben Vornamen *Barack* und *Hussein*, sowie denselben Familiennamen *Obama* aufweisen, ist die Wahrscheinlichkeit hoch, dass beide dieselbe Person repräsentieren. Die Namensübereinstimmung ist ein zentrales Kriterium für die Beurteilung der Identität der durch die verschiedenen Datenobjekte repräsentierten Person(en). Selbst eine perfekte Übereinstimmung ist aber keine Garantie dafür, das dieselbe Person gemeint ist. Dies kann man sich leicht bei gängigen Namen klar machen: Zehntausende von Personen heißen *John Smith*, *Muhammad Hussein*, *Hu Wong* oder *Hong Nguyen* – um nur ein paar wenige Beispiele häufiger Namen zu nennen.

Name Matching ist der Teil des Identity Matching, der sich mit der Übereinstimmung des Personennamens befasst. Personennamen sind im Identity Matching so zentral, dass gelegentlich sogar irreführenderweise Name Matching und Identity Matching synonym gebraucht werden. Doch werden im Identity Matching neben Personennamen noch weitere Merkmale zur Bestimmung der Übereinstimmung zweier Personendatenobjekte herangezogen, z.B. Geburtsdatum, Gründungsdatum, Geschlecht, Nationalität, Gründungsland, ID-Nummer, Steuernummer, Adresse.

Im Zusammenhang mit Identity Matching wird gelegentlich auch von *Identity Resolution* gesprochen. Beide Begriffe sind in weiten Teilen deckungsgleich und das meiste, was in diesem Buch über Identity Matching geschrieben steht, gilt auch für Identity Resolution. Der Unterschied beider Konzepte liegt darin, dass normalerweise Identity Resolution nicht für die Personensuche, sondern ausschließlich für den Personendatenabgleich verwendet wird. In diesem Sinne ist Identity Matching der allgemeinere Begriff.

B. Lisbach, *Linguistisches Identity Matching*, DOI 10.1007/978-3-8348-9791-6_2,

1.2 Datenprofile und Suchabfragen

In unserem Zusammenhang sind zwei Arten von Datenobjekten von besonderem Interesse: *Datenprofile* und *Suchabfragen*. Ohne auf technische Feinheiten einzugehen, sei ein Datenprofil als die strukturierte Menge jener Daten definiert, die über eine bestimmte Person in einer Datenbank oder Indexstruktur abgelegt sind. Man kann auch von einem *Personendatensatz* sprechen. Der Begriff Suchabfrage steht für die strukturierte Menge von Personendaten, mit deren Hilfe gesucht wird. Die Personendaten einer Suchabfrage werden gelegentlich auch als *Suchkriterien* oder *Suchattribute* bezeichnet.

Identity Matching vergleicht in der Regel entweder Suchabfragen mit Datenprofilen. Dann handelt es sich um eine *Personensuche*. Oder es werden Datenprofile untereinander verglichen. Wir sprechen dann von einem *Personendatenabgleich*.

Im Fall der Personensuche unterstützt Identity Matching die Beantwortung der Frage, ob sich von der Person, nach der gesucht wird, in einem bestimmten Datenpool ein Eintrag befindet. Im Fall des Personendatenabgleichs, also dem Vergleich von Datenprofilen untereinander, sollten zwei Anwendungsfälle unterschieden werden:

- Werden Datenprofile verglichen, die aus derselben inhaltlichen Klasse von Datenobjekten stammen, hat der Vergleich typischerweise die Funktion, Duplikate oder Dubletten zu erkennen und zu eliminieren. Das Ziel ist also die *Daten-Deduplizierung* (englisch: *Data Deduplication*).
- Werden hingegen Datenprofile aus Quellen unterschiedlicher inhaltlicher Klassen verglichen (z.B. Kundendaten mit Daten sanktionierter Parteien), dann hat der Datenabgleich die Funktion, Zusatzinformationen zu Personen, hier zu den Kunden, zu gewinnen. Konkret soll also für jeden Kunden ermittelt werden, ob er eine sanktionierte Partei darstellt, mit der Geschäftsbeziehungen zu unterhalten untersagt ist.

Abb. 1-1 zeigt einen typischen Anwendungsfall des Identity Matching in starker Vereinfachung zu dem Zwecke, die eben eingeführten Begriffe zu veranschaulichen: