



Worldmatch®

Internationale Daten abgleichen

Je mehr Kunden- und Lieferantenprozesse automatisiert werden, desto wichtiger werden korrekte, einheitliche Stammdaten. Während das Bewusstsein dafür einerseits zunimmt, wachsen andererseits die Anforderungen: Nach wie vor sind Datenabgleiche alles andere als trivial – und Fehler können teurer werden als je zuvor. Vor allem bei internationalen Abgleichen wird es kritisch: Das korrekte Verfahren ist ausschlaggebend.

Daten miteinander vergleichen

Immer wieder ist es notwendig, dass beispielsweise Stammdaten miteinander verglichen werden: etwa bei der Dublettenprüfung des eigenen Kundenstamms, beim Anlegen neuer Kundendaten, bei der Adress-Suche im Call-Center oder bei internationalen Adress-Abgleichen.

Im ersten Moment hört sich „Vergleichen“ einfach an. Doch was der Mensch beherrscht, fällt dem Computer schwer: „Gleich“ bedeutet für das elektronische Gehirn eine Übereinstimmung von 100 Prozent. Schon bei kleinsten Abweichungen wie einem Tippfehler versagt die Zuordnung. Der Computer versteht die Ähnlichkeit von „Mathias“ und „Matias“ nicht – zumindest nicht ohne programmierte Intelligenz. Für den Menschen dagegen sind beide Schreibweisen annähernd gleich. Genau auf diese Intelligenz kommt es bei elektronischen Vergleichen an.

Computer lernen Ähnlichkeit

Seit den sechziger Jahren bekämpfen clevere Programmierer die Doppeladressen. Zu Anfang war Rechenzeit noch teuer. So wurde der Flug von Apollo 11 im Jahr 1968 von einem Großcomputer berechnet, der in der Leistungsfähigkeit einem 286er-PC entsprach – mehr als die 50.000-fache (!) Rechenleistung steht heute auf vielen Schreibtischen! Um wirtschaftlich zu sein, mussten die ersten Verfahren also mit wenig Rechenzeit auskommen. Es war billiger, Verluste durch ein paar zu wenig gefundene Dubletten hinzunehmen, als einen größeren Computer anzuschaffen. Und so entstand das schnelle und ressourcensparende Matchcode-Verfahren.

Das Matchcode-Verfahren: schnell, aber ungenau

Statt bei zwei Datensätzen Buchstabe für Buchstabe zu vergleichen, werden mit ▶



dem Matchcode-Verfahren nur markante Punkte miteinander verglichen, beispielsweise PLZ, Hausnummer, der erste und der dritte Buchstabe des Nachnamens sowie der erste Buchstabe des Vornamens. Solange alle diese Dinge übereinstimmen, wird der Kunde als Dublette erkannt, trotz eventueller Abweichungen in den anderen Buchstaben.

Solche Verfahren benötigen kaum Rechenzeit, da dieser „Matchcode“ einfach gebildet wird und in einem Index für jede Adresse abgelegt werden kann – er muss also nicht jedes Mal für alle Adressen neu errechnet werden. Der Pferdefuß ist die schlechte Trennschärfe: „Maier“ zu „Meier“ wird zwar gefunden, aber schon bei „Meyer“ weicht der dritte, der relevante Buchstabe ab – der Matchcode versagt. Bei „Maier“ zu „Maihofer“ wird hingegen fälschlich eine Dublette gemeldet. Fazit: Das Matchcode-Verfahren ist veraltet.

Das phonetische Verfahren: Vertipper werden zum Verhängnis

Hier werden unterschiedliche Schreibweisen vereinheitlicht, indem ähnlich klingende Buchstaben in denselben Code verwandelt werden. P und B werden z.B. zur „1“, K und C und G zur „2“. Becker und Begger werden so als „gleich“ betrachtet.

Wie beim Matchcode kann auch hier der Phonetik-Code in einem Index gespeichert werden, was das Verfahren schnell macht. Es gibt eine Vielzahl solcher Phonetik-Verfahren, wobei die einfachen nur einzelne Buchstaben, die besseren auch

Buchstabengruppen („Sch“) betrachten. Das einfachste ist das Russell-Soundex-Verfahren, häufig nur Soundex genannt. Es ist weit verbreitet, macht aber auch sehr viele Fehler: So wird z.B. „Mehl“ und „Maier“ gleichgesetzt. Bei Dublettenabgleichen nennt man diese falschen Zuordnungen auch „Overkill“.

Eine moderne Phonetik kommt auch mit deutschen Umlauten klar und findet selbst kompliziertere Abweichungen wie „Kristof“ zu „Christoph“ oder „Klusoh“ zu „Cluseault“. Aber Achtung: Auch die beste Phonetik findet eben nur phonetische Fehler. Tippfehler oder Abkürzungen lassen sich damit nicht aufspüren. Deshalb sind phonetische Verfahren allein – zumindest für Firmenadressen – völlig unzureichend.

Unschärfe Ähnlichkeitsverfahren: die bessere Wahl

Da Computer immer leistungsfähiger und billiger geworden sind, ist es heute nicht mehr nötig, ein indizierbares Verfahren zu verwenden. Die Matchcode-Verfahren wurden daher von so genannten „unscharfen“ (engl. „fuzzy“) Verfahren abgelöst. Ein unscharfes Verfahren trifft keine Ja/Nein-Entscheidungen, sondern bestimmt den Grad der Ähnlichkeit. Durch Abwägen verschiedener Elemente der Adresse können so wesentlich höhere Trennschärfen erzielt werden: Sind beispielsweise Firmenname und Ansprechpartner zweier zu vergleichender Unternehmen sehr ähnlich, so kann trotz gänzlich anderer Straße (Umzug!) die ▶



**Carsten Kraus, Geschäftsführer
Omikron Data Quality GmbH**

Über die Omikron Data Quality GmbH

Alles begann 1992 mit der Entwicklung des neuartigen Ähnlichkeitsverfahrens für computergestützte Dublettenprüfung: FACT®. Schnell stellte sich heraus, dass die von Omikron geschaffene neue Algorithmik den bis dahin gebräuchlichen Matchcode- und Phonetik-Verfahren deutlich überlegen war. 1993 entstand daraus eine Abgleich-Software, die bis 1996 zu einer ganzheitlichen Software-Palette rund um Adressqualität ausgebaut wurde. 1997 und 2003 ersetzte jeweils eine neue Generation das komplette Portfolio. Dabei kamen neuste Programmiersprachen und Entwicklungsmethodiken zum Einsatz.

2007 wurde der Omikron Data Quality Server für den Einsatz in serviceorientierten Architekturen (SOA) fertiggestellt. Diese integrierbare Lösung versetzt Unternehmen in die Lage, alle in puncto Datenqualität kritischen Stellen einer IT-Landschaft abzusichern. Im selben Jahr setzte das neue Abgleich-Verfahren Worldmatch® für internationale Datenabgleiche völlig neue Maßstäbe.

Heute ist Omikron eines der führenden deutschen Unternehmen im Bereich Datenqualität. Für alle wichtigen Unternehmens-Anwendungen von SAP® bis Microsoft CRM® sind Integrationsmodule von Omikron erhältlich.

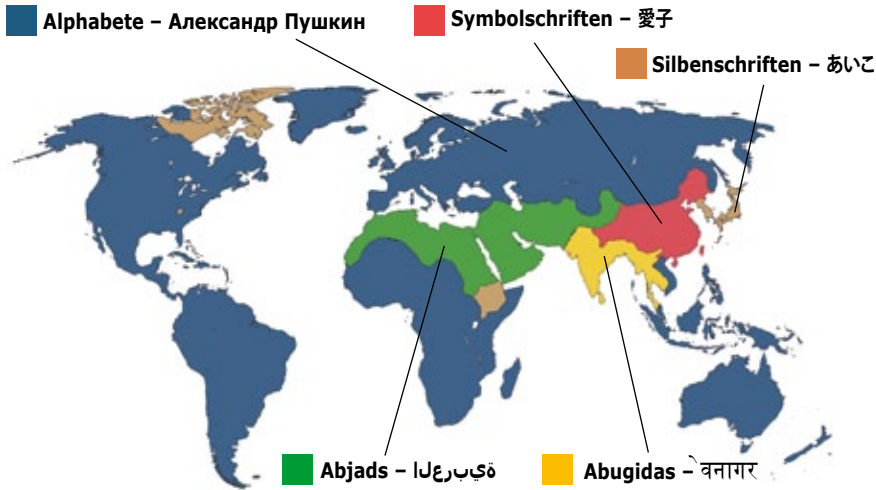


Abb.: Bild: Verteilung der Schriftsysteme Weltweit

Firma als Dublette zugeordnet werden. Bei annähernd identischer Straße und Hausnummer können andererseits stärkere Abweichungen im Firmennamen zugelassen werden. Diese Abwägung wäre etwa mit einem Matchcode-Verfahren nicht möglich.

Waren erste Verfahren wie Levenstein ausschließlich auf Tippfehler ausgerichtet, überwinden aktuelle Verfahren diese Hürde. So findet beispielsweise das FACT®-Verfahren alle oben aufgeführten Beispiele und kommt damit der menschlichen Fähigkeit, Ähnlichkeiten erkennen zu können, verdächtig nahe.

Datenabgleiche weltweit – die besondere Herausforderung

1993 erzielte Deutschland bereits 20 Prozent seines Brutto-Nationaleinkommens durch Exporte. Heute sind es mehr als 30 Prozent – und die Globalisierung nimmt weiter zu.

Durch das Zusammenwachsen der Wirtschaftskreisläufe sind auch die Anforderungen der Unternehmen an die eigenen Datenbestände enorm gestiegen: Firmensitz in Deutschland (lateinischer Zeichensatz), Niederlassung in Marokko (arabisch): Unternehmen, die international agieren, benötigen eine weltweit einsetzbare Datenbankstruktur und Prozesse, die ebenfalls weltweit funktionieren. Denn auch wenn die Logistik ein Zusammenrücken der Märkte ermöglicht, gibt es keine Hoffnung, dass sich die Welt in absehbarer Zeit auf eine gemeinsame

Standardsprache und -schrift einigen wird.

Und so erschwert eine Internationalisierung die Arbeit mit Daten, denn Abgleiche müssen nicht nur in der Muttersprache zuverlässig greifen, sondern auch über die eigene Grenze hinweg. Die Anforderungen an die Abgleich-Software steigen mit der Anzahl der Sprachen.

Ein Beispiel: Der irische Name Ewan hört sich korrekt ausgesprochen an wie „Juin“. Ein Sachbearbeiter in Deutschland, der nicht mit den sprachlichen Besonderheiten Irlands vertraut ist, wird den Namen vermutlich auch so in die Datenbank eintragen: Juin. Findet später ein Abgleich der deutschen Datenbank, der nur die Besonderheiten der deutschen Sprache berücksichtigt, mit den Daten der irischen Vertriebsgesellschaft statt, dann würde diese „Dublette“ nicht gefunden werden. Es ist also notwendig, dass die Abgleich-Software die Besonderheiten beider Sprachen kennt und miteinander in Beziehung setzen kann.

Solange man sich im lateinischen Schriftraum bewegt, kommt es lediglich auf einige Besonderheiten der Sprache an. Denkt man aber an die neuen Märkte in Russland, Indien und China, dann wird es wesentlich komplizierter. In diesen Sprachräumen wird mit völlig anderen Zeichensätzen geschrieben, deren Regelsätze sich zudem grundlegend unterscheiden. Im Arabischen wird von rechts nach links geschrieben und die Vokale entfallen. ▶

Zeichen der Welt

Alphabete – Александр

Bei Alphabeten entspricht jeder Buchstabe einem Sprachlaut. Man spricht auch von einer phonographischen Schrift.

Beispiele für Alphabete:

Lateinisch / Kyrillisch / Griechisch

Abjads – ٱبجدية

Abjads werden von rechts nach links geschrieben. Abjad ist eine reine Konsonantenschrift. Vokale werden bei den meisten Worten weggelassen, da sie für Einheimische offensichtlich sind und beim Sprechen einfach hinzugefügt werden.

Beispiele für Abjads:

Hebräisch / Arabisch

Abugidas – ॐनागर

Abugidas sind charakteristisch für die indischen und äthiopischen Schriften. Bei dieser Art von Schrift werden nur Konsonanten geschrieben. Es gibt Standardvokale, die verwendet werden. Kommt ein besonderer Vokal zum Einsatz, wird er mit einer speziellen Markierung gekennzeichnet. Abugidas werden auch als Zwischenstufe von Alphabet und Silbenschrift angesehen.

Beispiele für Abugidas:

Indisch (Devanagari) / Singhalesisch

Silbenschriften – あいこ

Silbenschriften gehören wie Alphabete auch zu den phonographischen Schriften. In einer Silbenschrift steht jedes Zeichen für eine Silbe.

Beispiele für Silbenschriften:

Japanisch (Hiragana) / Cherokee

Symbolschriften – 愛子

Bei Symbolschriften steht jedes Zeichen für ein komplettes Wort. Zusammengesetzte Wörter bestehen aus mehreren Symbolen. Symbolschriften werden auch als logographische Schrift bezeichnet.

Beispiele für Symbolschriften:

Chinesisch / Japanisch (Kanji)

Unicode Schwachstelle: Transkription

Unicode ist ein internationaler Standard, in dem für jedes sinntragende Schriftzeichen oder Textelement aller bekannten Schriftkulturen und Zeichensysteme ein digitaler Code feststeht. Mit einem einfachen Trick kann man so internationale Daten abgleichen. Dazu wird im ersten Schritt jedem Datensatz eine eindeutige Identifikationsnummer zugewiesen. Will man beispielsweise Kundendaten aus Deutschland und Japan miteinander abgleichen, könnte man die japanischen Zeichen ins lateinische Zeichensystem umwandeln (transliterieren). Der eigentliche Abgleich der Daten erfolgt dann mit lateinischen Zeichen.

Auf den ersten Blick macht diese Vorgehensweise einen soliden Eindruck. Bei genauerer Betrachtung lassen sich aber schnell die Schwächen erkennen. So gehen durch die Wandlung wichtige Informationen für einen unscharfen Abgleich verloren.

Das folgende Beispiel eines Vergleichs von deutschen und russischen Daten zeigt die Problematik: *Fyodor*, ein russischer Name, wird *Федор* geschrieben. Der kyrillische

Buchstabe e kann bei der Umwandlung ins Lateinische ebenfalls e oder zu ye werden. Ein diakritisches Zeichen über dem e – also *ë* – ändert die Betonung nochmals auf yo. Das Zeichen darf in der russischen Schrift auch ganz weggelassen werden.

Mit Unicode würde also je nach Regelsatz *Федор* (für *Fyodor*) entweder in *Fyedor* oder *Fedor* gewandelt werden. Beim anschließenden Abgleich mit den lateinischen Daten ergibt sich dann ein großes Problem, da die Ähnlichkeit zwischen *Fyodor* und *Fedor* eher gering ist. Eine zuverlässige Zuordnung ist so also nicht mehr möglich.

Fazit: „Unicode“ ist nicht genug! Transkription und Transliteration alleine funktionieren nicht, denn verschiedenen Schriften liegen unterschiedliche Funktionsprinzipien zu Grunde. Diese müssen beim Vergleich der Daten ebenfalls berücksichtigt werden.

Worldmatch® Sichere und verlässliche Abgleiche

Worldmatch ist ein Verfahren, das die internationalen Hürden meistert. Der

Vorteil dieses Verfahrens ist, dass die einzelnen Schriften nicht erst auf einen gemeinsamen Zeichensatz vereinheitlicht werden, sondern ein direkter Abgleich zwischen unterschiedlichen Zeichensätzen bzw. Schriften stattfindet.

Worldmatch transliteriert nicht, sondern assoziiert. Die Präzision der Abgleiche wird dadurch ungleich höher. Worldmatch prüft die verschiedenen Schriften gegeneinander, berücksichtigt die Besonderheiten der Sprachen und erkennt Ähnlichkeiten etwa bei Vertippern, Buchstabendrehern und Abkürzungen. So werden treffsichere internationale Abgleiche auf einem hohen Standard möglich.

Am Beispiel des russischen Namens „Fyodor“ (Федор) bedeutet das, dass Worldmatch alle möglichen lateinischen Schreibweisen beachtet. Also Fedor, Fyedor aber auch Fyodor.

Worldmatch steht sowohl als Funktion im Omikron Data Quality Server (Built for SOA) zur Verfügung und kann auch in Ihre eigenen Applikationen integriert werden.

Schreibweise	Zeichensatz
アイコ	Katakana
あいこ	Hiragana
あい子	Hiragana / Kanji
あ以子	Hiragana / Kanji
アイ子	Katakana / Kanji
あ衣子	Hiragana / Kanji
亜衣古	Kanji
亜伊子	Kanji
亜緯子	Kanji
亜以子	Kanji

Mögliche Schreibweisen des Namens: Aiko

Arabische Namen	
Vater	Hassan ibn Selim
Sohn	Yassir ibn Hassan

Im Arabischen ist der Name des Vaters Bestandteil der Kindernamen

Chinesische Namen	
张爱国	ZHANG Aiguo
张爱民	ZHANG Aimin
张爱党	ZHANG Aidang

Zhang = Familienname, Ai = Generationsname
Lediglich die letzte Silbe repräsentiert den eigentlichen Namen

Russische Namen	
Михаил Горбачёв	Michail Gorbatschow
Раиса Горбачёва	Raissa Gorbatschowa

Griechische Namen	
Πέτρος Κώτης	Petros Kotis
Αναστασία Κώτη	Anastasia Koti

Im Russischen wie auch im Griechischen ändert sich der Familienname je nach Geschlecht

Omikron Data Quality GmbH
Pfälzerstr. 35
75177 Pforzheim
Germany

Phone: +49 7231 12597 0
E-Mail: info@omikron.net
Web: www.omikron.net