

# Wie is wie in uw database?

De achtergronden van kennisgebaseerde oplossingen voor geautomatiseerde identificatie.

In vrijwel iedere organisatie, die werkt met omvangrijke databestanden, is het eenduidig en nauwkeurig herkennen van natuurlijke en/of rechtspersonen een probleem. Gegevens over personen en bedrijven vertegenwoordigen een grote commerciële waarde en zijn tegelijkertijd uitermate foutgevoelig.

Onjuist geschreven namen, verkeerde geslachtsindicaties, verschillende notatiemethoden voor een datum, niet herleidbare afkortingen en acroniemen, onvolledige en incorrecte adressen; het zijn slechts enkele voorbeelden van frequente fouten in grote databases. Het tegengaan en voorkomen van deze datavervuiling en haar gevolgen vereist meer dan traditionele matchingmethoden en waarschijnlijkheidscontroles. Het antwoord op de vraag "Wie is wie?" ligt besloten in identificatiesoftware, die landspecifieke kennis combineert met intelligente interpretatiemethodieken, gebaseerd op natuurlijke taalverwerking.

Deze editie van "In Detail" beschrijft de Nederlandse versie van HIquality® Identify, de identificatieoplossing van Human Inference.

## Beheren en beheersen van gegevens

Om een toegevoegde waarde te leveren aan alle bedrijfsprocessen, moeten gegevens actueel, correct, compleet en uniek zijn. Deze datakwaliteitsdimensies zijn de richtlijn voor het beheren en beheersen van data. Organisaties met substantiële klant- en prospectbestanden weten echter, dat de gemiddelde vervuilingsgraad van dergelijke bestanden tussen 10% en 15% procent ligt. Dit is wellicht geen al te hoog percentage in absolute zin, maar bij een bestand met 1.000.000 relaties is er wel degelijk sprake van een serieus datakwaliteitsprobleem.

Om de vervuiling van data te bestrijden, is het belangrijk om de oorzaken van de vervuiling te analyseren. Rechttoe, rechtaan zouden we kunnen zeggen dat de oorzaak ligt bij de mensen die met data werken: mensen maken immers fouten en niets menselijks is ons vreemd. Dat is echter een te simpele voorstelling van zaken. Aan de vervuiling van data ligt een complexer scala van oorzaken ten grondslag. Denk bijvoorbeeld aan het koppelen van bestanden, verouderde en/of inadequate bedrijfsprocessen en de invoer van gegevens via het internet.



Wanneer we de kwaliteit van de gegevens binnen de relevante processen willen beheren en beheersen, dan moeten we de gegevens kunnen identificeren.

## Identificatie begint bij interpretatie

Een goede identificatiemethode kan de gegevens van natuurlijke personen en rechtspersonen op een intelligente manier interpreteren. Daarbij moet onder andere rekening worden gehouden met de betekenis van woorden in een specifieke context, gebruiksnamen en vastlegging van namen in handelsregisters, afkortingen, synoniemen, acroniemen, fantasienamen, schrijffouten, notatiewijze van gegevens, standaards, normen en gelijkkluidendheid van woorden. Enkele voorbeelden:

|                       |                                   |
|-----------------------|-----------------------------------|
| Tuinaanleg Van Santen | Kon. Hoveniersbedr. Van Zanten BV |
| André Cauwberghs      | Andrée Koubergs                   |
| BMW                   | Bayerische Motorenwerke           |
| John van Dam          | Zhiang Van Tranh                  |
| Bakker De Koning      | J. Bakker-de Koning               |

HIQuality® Identify is een methode, die administratieve signalen (denk hierbij aan de analogie met fysieke signalen van personen: man, circa 1,80mtr. groot, bruin haar, bril dragend, etc.) aanmaakt van al deze kenmerken en ze vervolgens op een intelligente, fouttolerante manier in allerlei contexten identificeert.

Een administratief signaal kan bijvoorbeeld bestaan uit de volgende kenmerken (waarbij, waar nodig, zowel het woord als ook de klank van het woord wordt meegenomen):

|                |                |
|----------------|----------------|
| familienaam    | bedrijfsnaam   |
| voornaam       | rechtsvorm     |
| voorletters    | straat         |
| huisnummer     | postcode       |
| plaats         | geboortedatum  |
| telefoonnummer | geboorteplaats |

Afhankelijk van de context en het doel van de identificatie (bijvoorbeeld online zoeken of ontduubing) wordt de mate van overeenkomst tussen administratieve signalen vastgesteld ( zie hiervoor ook de paragraaf over toepassingsmogelijkheden).

### Kennis

De programmatuur moet dus signalen onderling vergelijken en daarbij rekening houden met:

- de betrouwbaarheid van de verschillende gegevens;
  - de waarschijnlijkheid van de waarde van de verschillende gegevens.
- Dit vereist uiteraard kennis over de verschillende kenmerken. Je moet weten wat je aan het vergelijken bent. We kennen allemaal de uitdrukking "appels met peren vergelijken". De benodigde kennis laat zich scheiden in twee categorieën:

- feiten
- kennis over de feiten

Om goed te kunnen vergelijken moet deze kennis alle ingevoerde woorden in alle betekenissen kennen. Waarom denken we dat het woord "Art" een voornaam is in Paul Simon & Art Garfunkel, een bedrijfswoord in Art Gallery Van der Vleut en een afkorting in Ton de Vos, Kookart. en Keukenstudio?

Paul Simon en **Art** Garfunkel  
**Art** Gallery Van der Vleut  
 Ton de Vos Kook**art**. en Keukenstudio

### Discriminatie waarde

Het zal duidelijk zijn dat vage signalen nauwelijks enige onderscheidende waarde hebben. Er kan hoogstens rekening gehouden worden met klank, spellingswijze en uniciteit in relatie tot de omvang van het bestand.

| Kenmerk     | inhoud1  | inhoud2   | inhoud3       |
|-------------|----------|-----------|---------------|
| naam        | Janssens | Jansen    | Jnasen        |
| voorletters | AB       | JPH       | P             |
| straat      |          |           | Dapperstraat  |
| huisnummer  |          |           | 10            |
| postcode    |          |           | 1017 GL       |
| plaats      | Sneek    | Amsterdam | A'dam         |
| geb.dat.    | 310763   | 030257    | 020357        |
| tel.        |          |           | 020 - 4562734 |
| geb.plaats  | Abcoude  |           | Abcoude       |



De naam Jansen heeft in Nederland nauwelijks enige discriminerende waarde. Naarmate meer gegevens bekend zijn (J.P.H.Jansen, Amsterdam, 030257) wordt de overeenkomst met enkele relaties groter en met andere minder. P. Jnsas, A'dam, 020357 kan de gezochte zijn en we kunnen meteen vaststellen dat A.B. Janssens, Sneek, 310763 onvoldoende overeenkomt.

Voorts moet de mate waarin wordt gediscrimineerd instelbaar zijn, met andere woorden: de mate waarin overeenkomst wordt aangetoond, moet instelbaar zijn. Bij het zoeken tijdens een telefoongesprek heb je nu eenmaal de beschikking over minder gegevens dan bijvoorbeeld tijdens het invoeren van een nieuwe relatie aan de hand van een aanmeldingsformulier. Het uitgangspunt kan ook anders zijn: bij het zoeken wil de beller graag gevonden worden, want hij of zij heeft bijvoorbeeld vragen over een levering, een factuur of een polis.

Bij een nieuwe aanmelding kan het zijn dat de aanmelder vindt dat er geen band meer is met het bedrijf. Bijvoorbeeld omdat hij of zij drie jaar geleden voor het laatst iets besteld heeft, inmiddels ergens anders woont en zichzelf daarom niet meer als klant beschouwt.

Of de aanmelder wil, om diverse redenen, bewust niet gevonden worden (postorderbedrijven, verzekeraars, enzovoort).

Binnen HIQuality Identify worden hiervoor wegingsfactoren gebruikt, waarmee de gebruiker (bijv. data stewards, database managers, etc.) de discriminatiewaarde van een administratief kenmerk kan instellen.

## Toepassingsmogelijkheden

HIQuality Identify biedt vele toepassingsmogelijkheden. Daarvan zijn zoeken, invoercontrole, vergelijken en ontdubbelen de meest gebruikte.

### 1. zoeken

Hiermee wordt over het algemeen online zoeken bedoeld. Door voor het opzoeken van relaties in de database gebruik te maken van intelligentie, kan de relatie met een beperkte set van invoergegevens gevonden. Doordat alles snel en zonder hinder wordt gevonden, zal de gebruiker niet in de verleiding komen het zoeken te staken en de relatie "gewoon als nieuwe" op te voeren. In een callcenter bijvoorbeeld is snelheid en accuratesse in het zoekproces van groot belang.

### 2. invoercontrole

Invoercontrole is een specifieke toepassing van online zoeken. Het komt erop neer dat bij het vastleggen van een nieuwe relatie het computersysteem achter de schermen bepaalt of deze al mogelijk voorkomt. Enkele bedrijven passen een vorm van invoercontrole toe. De meeste hanteren het principe "eerst zoeken niet vinden opvoeren".

Bij invoercontrole wordt in eerste instantie gedaan alsof de relatie nieuw is.

In tegenstelling tot zoeken worden nu alle gegevens ingevuld. Hierdoor is de discriminatiewaarde groter. Met andere woorden: er zijn slechts weinig bestaande relaties die veel lijken op de nieuwe. Dit impliceert dat invoercontrole alleen zinvol is als het relatiebestand minder dan de helft van het universum omvat, anders wordt te vaak een dubbele gesignaleerd en dan weegt alles invoeren niet op tegen het eerst gewoon zoeken.

### 3. bestandsvergelijking (matchen, crossen, enzovoort)

Er zijn veel situaties waarin gegevens moeten worden geïmporteerd in het eigen bestand. Denk hierbij aan fusies, acquisities of verrijking van bestanden met referentiedata. Onder andere door verschillen in adressering en schrijfwijze van de naam kunnen doublures ontstaan. In dat geval is het verstandig om tijdens het importeren de toegevoerde gegevens meteen op intelligente wijze te vergelijken met de inhoud van het eigen relatiebestand.

Komt het importeren van bestanden sporadisch voor, dan is het wellicht handig het aangekochte bestand zonder meer te importeren in het relatiebestand en daarna het geheel te ontdubbelen.

Het toepassen van intelligentie bij bestandsvergelijking is niet alleen noodzaak bij het importeren van gegevens, maar ook voor toepassingen als het maken van een doorsnede uit twee bestanden, het matchen tegen allerlei suspect-lists in het kader van compliance, het aggregeren van meerdere bestanden naar een centraal bestand en het vergelijken ten behoeve van een inzicht in de markt (marktpenetratie-analyse).

### 4. ontdubbelen

De voorgaande drie toepassingsgebieden zorgen ervoor dat vervuiling wordt voorkomen aan de bron. Ontdubbelen is dan in principe alleen nodig als een intelligent beheersmechanisme wordt ingebouwd in een bestaand systeem met een reeds

gevuld relatiebestand. Niettemin kan periodiek ontdubbelen zinvol zijn, want bij zoeken en invoercontrole kan men nooit voorkomen dat een gebruiker een relatie toch een tweede keer opvoert. De frequentie van het ontdubbelen is mede afhankelijk van de omvang van het bestand, de mutatiegraad en de zorg waarmee gewerkt wordt. Bovendien zal de lijst met dubbeln niet zo groot zijn: alleen zinvolle overeenkomsten worden gemeld. Met de gebruikelijke mutatieprocedure kan dan corrigerend worden opgetreden.

## Conclusie

Een goede identificatiemethode is een absolute voorwaarde voor correcte en efficiënte bedrijfsprocessen. Foute of onvolledige identificatie leidt onder andere tot ernstige bestandsvervuiling, een lage analysebetrouwbaarheid (geen centraal klantbeeld), kostenstijgingen en imagoschade. HIQuality Identify is een methode die gebaseerd is op natuurlijke taalverwerking en fuzzy-logic-technieken. Hierdoor onderscheidt HIQuality Identify zich van alle andere traditionele identificatieoplossingen. Zowel voor online verwerking als batchverwerking wordt uitgegaan van administratieve signalen, waarin alle relevante gegevens zijn opgenomen. Een geavanceerde fonologiemodule zorgt voor een fouttolerante adaptatie van klankgelijkheid en klankverschillen. Landspecifieke kennis corpora geven de methode de munitie voor interpretatie volgens de principes van natuurlijke taalverwerking. Daarnaast beschikt HIQuality Identify over krachtige contextanalyse-algoritmen om uit namen en adressen de significante en identificerende gegevens te bepalen. Door de instelling van wegingsfactoren in combinatie met diverse intelligente evaluatieslagen, is de methode volledig configureerbaar. Daarom geeft HIQuality Identify u volledige controle op alle identificatieproblemen.

**Human Inference is marktleider op het gebied van oplossingen voor datakwaliteitsmanagement. Onze oplossingen zijn gebaseerd op een geavanceerde taal- en cultuurspecifieke aanpak, die het mogelijk maakt om klantgegevens intelligent te beheren en een eenduidig klantbeeld te genereren. De klanten van Human Inference gebruiken onze producten voor verbeterde customer intimacy, operational excellence en succesvolle oplossingen op het gebied van compliance.**