

## Dataquality hoezo?

*Auteur: Patric Schoenaker Consultant bij QNH  
Enterprise Intelligence.*

*Co auteur: Paul Houben Projectleider & Consultant bij  
QNH Euregio. [www.qnh.eu](http://www.qnh.eu)*



*Niets uit dit document mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand of openbaar gemaakt in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen of op enig andere manier, zonder voorafgaande schriftelijke toestemming van QNH.*

### **Dataquality.... Hoezo?**

Dataquality wordt binnen ondernemingen vaak als iets moeilijks gezien. Het is complex, tijdrovend, en algemeen moeilijk te realiseren. Vaak speelt ook de gedachte bij de organisatie dat het ook weer niet zo belangrijk is.

Voor diverse businessprocessen is dataquality echter van doorslaggevend belang. Vaak worden BI (Business Intelligence) trajecten gestart om deze businessprocessen te ondersteunen en van juiste, volledige en tijdige stuurinformatie te voorzien. Wat is echter de waarde van de rapportages die uit dat BI traject komen, als de onderliggende gegevens onbetrouwbaar zijn en/of dubbel zijn geteld? Hoe kan men dan de juiste beslissingen nemen? Hoe kan het BI traject werkelijk helpen en inzicht bieden, als de gegevens waarop het gebaseerd is niet juist zijn? In marketing en sales ziet men bijvoorbeeld een steeds groter wordende focus op het bereiken van Consumer Intimacy. Dit is een bedrijfsdoelstelling die zonder goede dataquality simpelweg niet te bereiken is.

In de volgende uiteenzetting zal in het kort Consumer Intimacy en het belang van dataquality voor het bereiken daarvan worden uiteengezet. Als laatste wordt beschreven hoe Trillium Software System, een softwarepakket specifiek voor het oplossen van dataquality kwesties, kan helpen bij het bereiken van deze hogere datakwaliteit.

### **Consumer Intimacy**

Wat is het belang van Consumer Intimacy voor ondernemingen?

Door consumenten beter aan zich te binden verwachten ondernemingen meer omzet te realiseren, en producten en/of diensten te kunnen leveren die beter aansluiten bij de verwachtingspatronen van deze consumenten. Consumer Intimacy houdt dus in dat beter wordt begrepen hoe de consument denkt, wat hij wil, en wat hij verwacht.

Het verzamelen van zoveel mogelijk consumentengegevens is vervolgens één stap in het proces om uiteindelijk consumer intimacy te bereiken. Als een grote verzameling consumentengegevens is aangelegd, kan deze verzameling worden gesegmenteerd en zodoende inzicht worden verkregen in diverse eigenschappen van deze consumenten.

Door het verkregen inzicht kunnen campagnes beter worden gebudgetteerd en zullen vervolgens resulteren in hogere af- en omzet. Daarnaast worden kostenbesparingen gerealiseerd doordat minder geld wordt verspild aan onnodige campagnes. Of campagnes die niet de gewenste doelgroep bereiken. Producten kunnen beter worden toegespitst op de potentiële afnemers. Dit doordat bekend is welke groepen van consumenten geïnteresseerd zijn in welke producten, of welke functionaliteiten van producten.

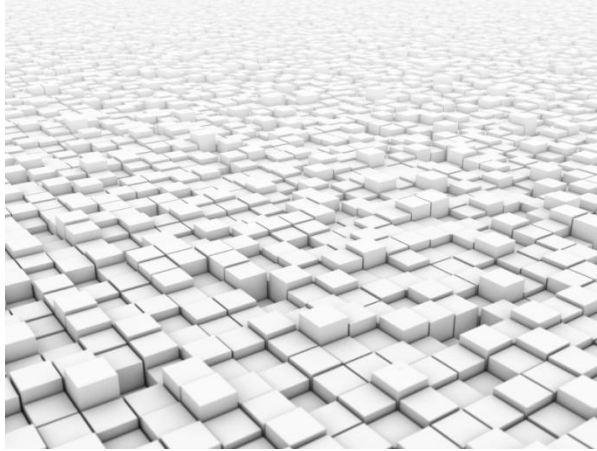
Wat denkt een potentiële klant namelijk als het omgekeerde gebeurt?

Als hij telkens wordt benaderd met campagnes die niet op hem van toepassing zijn? Als hij telkens wordt benaderd met acties voor producten die hij net heeft aangeschaft? Als hij te vaak wordt benaderd? Als hij telkens wordt benaderd met producten waarvan de prijs ver buiten zijn budget ligt? Of als hij 3 keer wordt benaderd met dezelfde campagne?

Het resultaat zal zijn dat de klant aangeeft niet meer benaderd te willen worden, of alle email campagnes zullen in zijn "ongewenste email" map belanden. Daarnaast is er een grote kans dat via 'word-of-mouth' het bedrijfsimago zal verslechteren.

Goed gerichte campagnes zullen de consument echter aansporen, enthousiasmeren en uiteindelijk kan de consument veranderen in een zogenaamde promotor van de organisatie. Het aantal verkochte producten zal toenemen. Daarbij zal via 'word-of-mouth' reclame het bedrijfsimago verbeteren.

Nu de grote verzameling consumentengegevens is aangelegd stellen we de vragen: Zijn we er dan? Kunnen we nu aan de slag? Worden campagnes nu beter gebudgetteerd? Is nu bekend wat de consument van ons als onderneming verwacht? Is er nu goed inzicht in onze consumentenbase?



Zie figuur links een grote hoeveelheid gegevenselementen ongeordend met verschillende kwaliteitsniveaus.

Het antwoord op deze vragen is helaas nog “nee”.  
**Want wat is eigenlijk de kwaliteit van deze consumentengegevens?**

De kwaliteit van de gegevens waar campagnes en analyses op gebaseerd worden is van doorslaggevend belang. Slechte gegevenskwaliteit leidt tot verkeerde inzichten en inschattingen, en dus verkeerde (management-)beslissingen.

Als een consumentendatabase bijvoorbeeld veel dubblures bevat zal informatie over de grootte van doelgroepen bijvoorbeeld nooit accuraat kunnen zijn. Ook verdelingen over die doelgroepen (leeftijdscategorieën, geslachten, inkomensverdelingen, etc.) zullen nooit kunnen kloppen!

Correctheid van de individuele gegevenselementen zelf is natuurlijk ook van belang. Zonder juiste geslachtsinformatie kunnen consumenten nooit correct aangesproken worden en kunnen specifieke campagnes nooit gericht worden op alleen vrouwen of mannen. Bijvoorbeeld als men een scheerapparaat of ladyshave wil promoten. Zonder correcte adresgegevens zijn demografische analyses niet mogelijk. (althans: niet met een bruikbaar resultaat). Daarnaast zal een groot deel van het verstuurd materiaal niet bij de beoogde doelgroep aankomen. Men zal moeten voorkomen babyvoeding te slijten aan kinderloze paren. Dure horloges aan mensen met een bijstandsuitkering. Of dat mensen voor de verontreinigheffing worden aangeslagen als gezin terwijl ze alleenstaand zijn!

De effecten hiervan zijn natuurlijk wederom legio. Minder resultaat van uw campagne, potentiële schade aan uw ‘brand image’ en verspilde budgetten aan onnodige direct mail zijn slechts enkele voorbeelden.

Bovenstaande uiteenzetting heeft echter vooral betrekking op zogenaamde NAW-gegevens (Naam, Adres, Woonplaats en andere contactinformatie). Slechte kwaliteit van bedrijfseigen gegevens heeft echter evenzeer een grote impact op de performance van de onderneming. Is bekend welke consumenten welke producten afnemen? En zitten daar dubbeltellingen in? Worden in rapportages altijd dezelfde meeteenheden gebruikt? Worden in de diverse systemen alleen product- en type aanduidingen geregistreerd die ook daadwerkelijk bestaan en geleverd worden?

Samenvattend kunnen we stellen dat de kwaliteit van de gegevens waarop analyses, rapportages en campagnes worden gebaseerd van doorslaggevend belang is. Dit voor het succes van een onderneming, en de wijze waarop deze wordt gemanaged.

### ***Het verhogen van het kwaliteitsniveau.***

Laten we nu eens kijken naar wat we moeten doen om de gegevenskwaliteit omhoog te krijgen. Gegevens zullen moeten worden herkend, gecorrigeerd, aangevuld en ontdebeld (of “gededuplicateerd”).

Het mag duidelijk zijn dat dit geen handmatige actie is. Het gaat in sommige gevallen om miljoenen consumenten, die elk meerdere adressen, telefoonnummers, email-adressen en producten kunnen hebben. Handmatige schoonmaakacties zijn tijdrovend daardoor duur, slecht beheer(s)baar, en niet integreerbaar met (online) applicaties. Bovendien zal dit schoonmaken (en schoonhouden!) een telkens terugkerend proces zijn. Het klantenbestand is immers ook niet statisch!

“Geautomatiseerde handmatige” schoonmaakacties bijvoorbeeld in maatwerkscripts bieden ook geen soelaas. De grote hoeveelheid aan mogelijke afwijkingen in beschikbare gegevens houdt in dat de cleaning oplossing continue zal moeten worden aangepast. Daarnaast bieden dit soort oplossingen ook nooit de flexibiliteit, performance, kwaliteit en integreerbaarheid die in de markt beschikbare dataquality oplossingen bieden.

Men zal dus over moeten gaan op implementatie van een goed performende, hoge kwaliteit biedende, flexibele en goed instelbare dataquality oplossing. Hiermee kan de kwaliteit van beschikbare gegevens structureel en uniform verbeterd worden, zonder oeverloze exercities van handmatige datamanipulatie.

***Wanneer moeten deze schoningsacties dan plaatsvinden?*** Of op welke plaats in het proces moet deze dataquality oplossing dan worden ingezet?

Idealiter wordt ervoor gezorgd dat ‘vuile’ gegevens niet in de diverse systemen terecht kunnen komen. Het moet zo moeilijk mogelijk zijn deze ‘vuile’ gegevens in te voeren. Dit kan bijvoorbeeld gerealiseerd worden door dataquality oplossingen in te zetten op de verschillende punten waar informatie wordt ingevoerd. Hierdoor kan ervoor gezorgd worden dat gegevens “schoon” worden aangemaakt, en we dus bijvoorbeeld alleen echt bestaande en correct gespelde straatnamen opslaan, volgens in een bepaald land geldende normen. Ook kunnen deze dataquality oplossingen ervoor zorgen dat er deduplicatie plaatsvindt; in het geval klantgegevens worden ingevoerd die al voorkomen in de betreffende systemen, kan ervoor gezorgd worden dat deze niet nogmaals worden opgeslagen.

In sommige gevallen worden echter ook klantgegevens aangeleverd door derden zoals marketing of reclamebureaus. Om ook deze gegevens te kunnen corrigeren en dedupliceren zullen hiervoor voorzieningen moeten worden ingericht. In de markt verkrijgbare dataquality oplossingen bieden voorzieningen om gegevens in “batch” te verwerken, zodat ze na aanlevering geschoond en gededupliceerd kunnen worden voor ze aan databases of datawarehouses worden toegevoegd. Sommige van de geboden oplossingen zijn in meer of mindere mate geïntegreerd met bestaande ETL oplossingen, zodat de gegevens al tijdens het inlaadproces geschoond worden.

Beide scenario's zullen moeten worden geïmplementeerd om uiteindelijk te komen tot een consumentenbase die van een dusdanig kwaliteitsniveau is. Met het doel dat de resultaten die hierboven beschreven staan behaald kunnen worden.



Zie bovenstaand figuur het toevoegen van waarde door ordening en het verhogen van kwaliteit

## **DQ oplossing met Trillium Software System**

Trillium Software System versie 11.5 ("TSS") is een van de dataquality oplossingen die voorziet in alle eerder genoemde eisen die je aan een DQ tool kunt stellen.

Door het inzetten van een DQ product als Trillium Software System heeft een onderneming de juiste gereedschappen in huis om bestaande en nieuwe data te onderzoeken (profilen), op te schonen, aan te vullen en te dedupliceren. (zowel bij data-entry als in batch). Met TSS is deze functionaliteit beschikbaar in één geïntegreerde suite.

### **Profiling**

Op diverse momenten tijdens de uitvoering van een project zal het bijvoorbeeld nodig zijn de 'stand van zaken' op te nemen. Door middel van de profiling functionaliteit (Trillium Software Discovery) kan snel en eenvoudig naar de verschillende aspecten van de (bedrijfs-)data worden gekeken. Al tijdens het inladen van data in Discovery wordt deze geanalyseerd waardoor tijdens het 'profilen' geen langdurige (database) queries meer nodig zijn. Dit houdt in dat informatie over uniciteit, spreiding, semaphones, soundexes, null-counts, patronen, lengtes van data-elementen en de validatie van businessrules al tijdens het inladen wordt gegenereerd. Toch is het laadproces zeer snel want miljoenen rijen klantgegevens kunnen in minuten worden ingeladen.

Evenmin hoeft vooraf bedacht te worden welke aspecten van de data men wil onderzoeken. Een niet te onderschatten eigenschap van een DQ tool!

Als namelijk van te voren al bekend is welke problemen er waarschijnlijk zullen opduiken is het niet nodig te profileren, maar kunnen die problemen meteen aangepakt worden.

Het belang van het profileren van data is juist problemen op te sporen die wellicht minder voor de hand liggen, maar toch in grote mate de kwaliteit van de gegevensverzameling bepalen. Zo is het bijvoorbeeld lastig vooraf te bedenken dat door een fout in de laadprocedures bij alle Italiaanse consumenten de laatste 3 letters van de straatnaam vóór aan de straatnaam zijn 'geplakt', zodat "via XX Settembre" wordt opgeslagen als "brevia XX Settem"!

Nadat de data geprofiled is moeten de geconstateerde problemen natuurlijk opgelost worden. Zoals eerder vermeld moet dit idealiter op twee plaatsen gebeuren. Bij data-entry en voordat de gegevens aan de database worden toegevoegd. Voor beide situaties is er bij TSS een geïntegreerde oplossing beschikbaar.

### **Cleaning & Deduplicatie**

Allereerst zal bij data-entry ervoor gezorgd moeten worden dat er geen 'vuile' gegevens meer aan de systemen worden toegevoegd. Anders is het spreekwoordelijk gezegd dweilen met de kraan open. TSS biedt hiervoor een standaard webservice die vanuit elke (web-)applicatie aangeroepen kan worden. Hierdoor worden gegevens al tijdens het invoerproces geschoond en al (deels) gededupliceerd. Indien een consument wordt ingevoerd en gelijksoortige consumenten in de database worden gevonden (op basis van naam, straat en plaats bijvoorbeeld) dit is echter compleet configureerbaar, wordt een pop-up getoond waaruit de medewerker de juiste consument kan kiezen. Dit werkt ook wanneer data niet in de juiste volgorde, of zelfs in de juiste attributen wordt ingevoerd! (achternaam in straatnaamattribuut, voorletters in voornaamattribuut). Deze functionaliteit van TSS is inmiddels gecertificeerd en integreerbaar in CRM en ERP systemen van bekende leveranciers zoals Oracle (Siebel) en SAP.

Daarnaast is het wenselijk data in 'batch' te kunnen cleanen en dedupliceren. Ook hier biedt TSS volledig geïntegreerde mogelijkheden, die naadloos samenwerken met de hierboven beschreven webservice en Profiling functionaliteiten.

De volgende stappen in het proces worden uitgevoerd door middel van Trillium Quality. Dit is het krachtigste onderdeel van de suite waarmee gegevens geschoond, verrijkt en gededupliceerd kunnen worden.

### **Standaardiseren**

De eerste fase van dit proces is het standaardiseren van de gegevens. TSS Quality herkent op basis van zeer uitgebreide bibliotheken en patroonherkenning diverse gegevenselementen (namen, adressen, telefoonnummers, e-mailadressen, bedrijfsgegevens etc.) en zet deze in de juiste attributen. Deze attributen worden vervolgens geschoond op basis van dezelfde bibliotheken. De bibliotheken worden door Trillium verkregen bij toonaangevende instanties in diverse landen, en worden onderhouden en verspreid onder de afnemers van haar product. Bedrijfseigen gegevens kunnen eenvoudig aan deze bibliotheken worden toegevoegd.

### **Verrijken**

Na het standaardiseren worden gegevens waar mogelijk verrijkt. Op basis van bepaalde gegevenselementen kunnen andere gegevenselementen worden afgeleid. Zo kan bijvoorbeeld vaak op basis van de straatnaam een postcode worden afgeleid (in sommige landen ook andersom), of kan op basis van een voornaam een geslacht worden afgeleid.

### **Dedupliceren**

Na het verrijken kunnen gegevens worden gededupliceerd. Gestandaardiseerde en verrijkte gegevenselementen (correcte straatnamen, volledige postcodes, achternamen zonder initialen of voornamen) kunnen met elkaar worden vergeleken op basis van volledig configureerbare parameters. Zo kan worden gedefinieerd:

- Op basis van welke elementen vergeleken moet worden.
- Hoe deze elementen vergeleken moeten worden. (gewone vergelijking of door middel van een speciale 'fuzzy' functie, zoals de "business name" functie)
- Hoe zwaar de vergelijking weegt in combinatie met andere vergelijkingen. (gewicht)
- Bij welke kwaliteit (combinatie van gewichten van verschillende vergelijkingen) de records als een "pass" een "suspect" of een "fail" worden gezien.

Het gevolg van deze aanpak is dat deduplicatie niet gebaseerd is op waarschijnlijkheden, maar eerder gesproken kan worden van 'precision matching'. Matching kan zeer specifiek afgestemd worden op de behoeften van iedere individuele organisatie en omstandigheid.

Aan het einde van het deduplicatieproces kan een zogenaamde "survivor" worden bepaald. Een record dat na deduplicatie overblijft. Deze survivor kan worden gekozen op basis van de oorspronkelijke records (met geschoonde en gestandaardiseerde gegevenselementen). Men kan echter ook een survivor samenstellen uit de gegevenselementen van diverse records.

Bijvoorbeeld: "Neem de adresgegevens uit het record met de hoogste adreskwaliteit, de productnummers uit het record met de meest recente datum, het email adres wat het meest recent gebruikt is en het mobiele telefoonnummer indien aanwezig, anders het vaste telefoonnummer".

### **International handling**

Het is goed om te weten dat TSS volledig UTF-8 compliant is. Dit wil zeggen dat bovenstaande werking ook op gaat voor landen met afwijkende character sets zoals China, Griekenland, Israël (Hebreeuws) en Korea. Het duiden van de resultaten van TSS Discovery en van het TSS Quality proces zullen wat praktische problemen kennen. Ten slotte is niet iedereen het Mandarijn machtig. De standaard functionaliteit van TSS Quality is echter zodanig sterk dat ook al grote kwaliteitsverbeteringen kunnen worden behaald zonder dat uitvoering te finetunen nodig is.

### **Aanbeveling.**

Het inzetten van DQ tooling als TSS alleen is niet voldoende om hoge kwaliteit van beschikbare gegevens in een organisatie te blijven waarborgen. DQ zal een integraal aspect van de organisatie moeten uitmaken. Men zal (kwaliteits-)standaarden moeten vaststellen en deze moeten vastleggen, onderhouden en implementeren. Referentiegegevens zullen moeten worden onderhouden en centraal worden vastgelegd. De organisatie zal zich moeten conformeren aan deze nieuwe standaarden en zal ermee moeten (leren) werken.

**Conclusie**

Voor het bereiken van Consumer Intimacy en het toepassen van Marketing Intelligence is het beschikken over hoge-kwaliteit data onontbeerlijk. Lage datakwaliteit zal leiden tot misverstanden bij strategische, tactische en operationele analyses. Slecht getargette marketing campagnes en verminderd inzicht in de behoeften, voorkeuren, houding en gedrag van de consument. Door data bij point-of-entry, en vóór opslag te laten schonen en ontdebelen door een goede DQ tool, kan dit worden voorkomen.

Trillium Software System is ook volgens Gartner, die hen als leider in het Magic Quadrant ziet, een van de beste oplossingen voor het bereiken van een hoger niveau van datakwaliteit. De geïntegreerde suite van Trillium levert een volledig configureerbare snelle, accurate en in uw systemen integreerbare oplossing van hoog niveau. Er zijn veel processen binnen een organisatie waar naar gekeken moet worden om integraal een hoger datakwaliteitsniveau te bereiken.