Informatiekwaliteit? Erover praten helpt.

Peter Alons (Dr. P.W.F. Alons)
en
Rob Arntz (Drs. Ing. R.G. Arntz)

Atos Origin/Business Intelligence-CRM

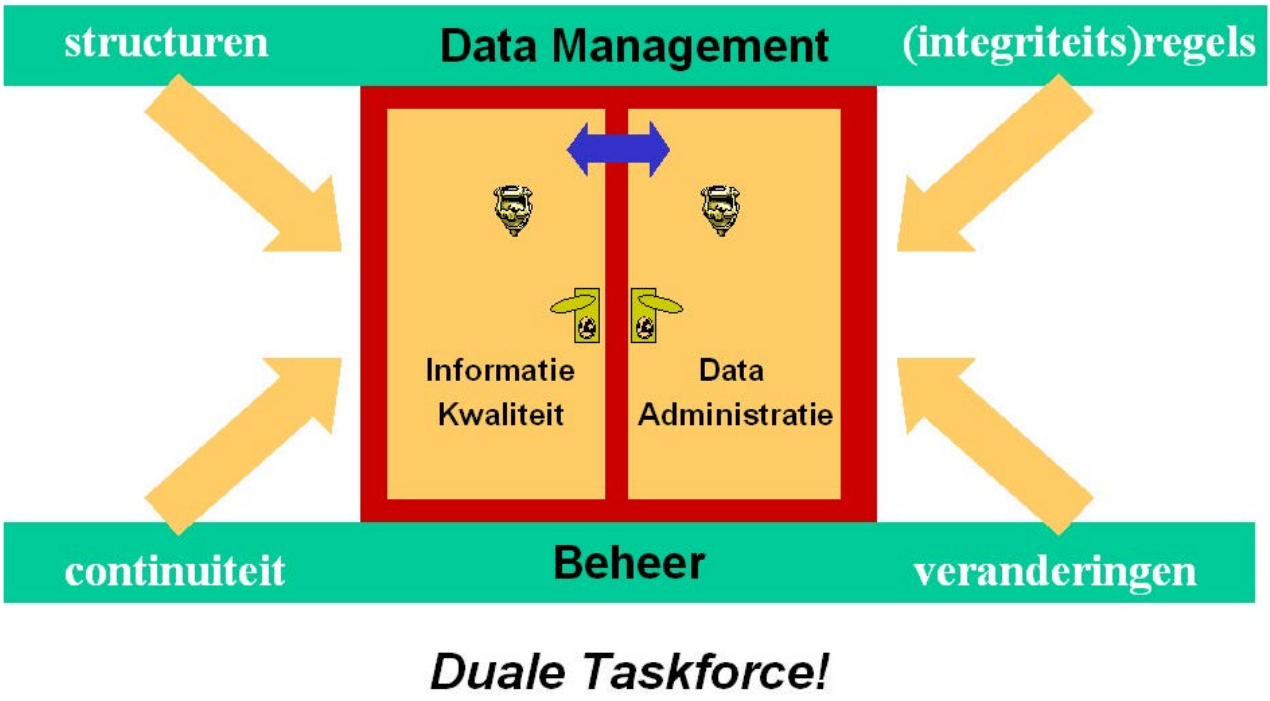
In een vorig artikel in DB/M [1] hebben wij de begrippen datakwaliteit en informatiekwaliteit besproken. We betoogden, dat er een enorm verschil bestaat tussen die twee. Datakwaliteit is al gauw een wollig begrip, waaronder door verschillende IT-specialisten verschillende dingen wordt verstaan. Informatiekwaliteit is een begrip waarvan door eindgebruikers altijd een helder beeld valt te geven: de informatie deugt of hij deugt niet. Men zou kunnen stellen, dat datakwaliteit an sich een *ding* is, maar goede informatiekwaliteit een tastbaar *feit*. Dat brengt ons bij een ander artikel van ons in DB/M [2], waarin wij het verschil tussen dingen en feiten hebben uiteengezet. In dit artikel brengen we deze twee gedachtegangen bij elkaar. De uitkomst daarvan doet volledig recht aan alle uitgangspunten en doelstellingen van onze Metadata Frame aanpak, die besproken zijn in een tweetal artikelen in DB/M [3, 4] over de weg naar betrouwbaar en betaalbaar Metadata Management.

Om helder te maken hoe goede informatie-kwaliteit daadwerkelijk geëffectueerd kan worden, geven we eerst weer twee voorbeelden van de verschrikkelijke kosten van non-kwaliteit in informatie. Het eerste staat ook in ons vorige artikel [1]. Het tweede is één de van voorbeelden, die wij sindsdien in het nieuws hebben aangetroffen.

1. [Teletekst, 2002]: In 2002 werd in een ziekenhuis in Amerika de bloedgroep bij een donorhart foutief geregistreerd als A+. Kort daarna werd het hart geïmplanteerd bij een jonge vrouw. Toen bleek, dat de bloedtypering niet klopte, was het te laat: een nieuwe operatie was niet mogelijk en de vrouw overleed kort daarna.
2. [TOKIO, december 2005] - Een medewerker van het effectenkantoor van de Japanse bank Mizuho Financial Group heeft waarschijnlijk de duurste typefout ooit gemaakt. Door op een verkeerde knop te drukken, werden voor 1 yen per stuk 610.000 aandelen in het Japanse bedrijf J-Com verkocht. Dit had eigenlijk een opdracht voor de verkoop van één aandeel voor 610.000 yen moeten zijn. De fout kostte de bank 27 miljard yen (190,5 miljoen euro). De president van Mizuho Securities, Makoto Fukuda, moest de volgende avond diep buigen. Hij bood de investeerders, de opdrachtgever en alle anderen zijn excuses aan.

De reactie over deze zaak van een analist op www.bloomber.com was als volgt:
"It's absurd that the exchange doesn't have any system to reject such incomprehensible orders," said Takao Saga, a senior economist at Japan Securities Research Institute. "The exchange should have at least halted trades on J-Com yesterday when officials noticed there was an incomprehensible order. Then, they could have checked what was going on, and canceled the order, which was made by mistake."

Er zijn natuurlijk verschillende manieren om dit soort problemen van geval tot geval te *verhelpen*. Het management van Mizuho Financial Group heeft voor hun situatie zelf al een aanpak gepubliceerd. Wij willen hier graag een algemeen bruikbare aanpak bespreken, die ertoe bijdraagt om dit soort fouten te *voorkomen* in plaats van te verhelpen, zoals in bovenstaande standaard reactie wordt gesuggereerd. Het is een oplossing die bij gebruik van onze Metadata Frame methode niet moeilijk is om te realiseren. In ons vorig artikel [1] toonden we een figuur met twee wel te onderscheiden deuren naar functies binnen een bedrijf: een deur met Informatiekwaliteit en een deur met Data Administratie (zie figuur 1). Onze oplossing komt vanachter de rechterdeur, maar hij dient volledig de linkerdeur, waarachter zoals de voorbeelden aantonen in elk bedrijf veel geld wordt verdiend of verloren.



Figuur 1: ***De twee deuren naar functies binnen een bedrijf***

Hoe kunnen we informatiekwaliteit van achter de rechterdeur vandaan tot een tastbaar feit helpen maken? We moeten daartoe in elk geval recht doen aan de conclusie van ons vorige artikel: dat kwaliteit en dus ook informatiekwaliteit in de eerste plaats een zaak van mensen is en hun gedrag. Technische hoogstandjes achter de rechterdeur zijn weinig effectief, als ze niet het juiste gedrag van gebruikers stimuleren. Dit laatste doen we door consequent gebruik te maken van communicatie.

Voor goede datakwaliteit is in de eerste plaats een valide datamodel nodig. Immers, als de structuur van de data niet goed is, kan een computer gemakkelijk verkeerde feiten opslaan zonder dat hijzelf of een menselijke gebruiker dat gebeuren kan vaststellen. Binnen het Metadata Frame gebruiken we FCO-IM [5] als methode om op basis van communicatie met domeindeskundigen een volledig gevalideerd datamodel model op te stellen. Via deze communicatie leiden we niet alleen een volledig correcte structuur in de zin van tabellen en kolommen af: ook de beperkingsregels worden door middel van communicatie opgespoord en vastgelegd.

In een vereenvoudigd voorbeeld van een informatiemodel dat gebruikt kan worden in de medische wereld bij de opname van een patiënt op de eerste hulp wordt via een formulier onder andere de volgende feiten vastgelegd:

Patiëntnummer:	**3427839**
Naam:	**Pieter Jansen**
Woonplaats:	**Breda**
Bloedgroep:	**A+**

De zinnen die hierbij uitgesproken kunnen worden, zijn:

“Patiënt 3427839 heeft voornaam Pieter”
“Patiënt 3427839 heeft achternaam Jansen”
“Patiënt 3427839 woont in Breda”
“Patiënt 3427839 heeft bloedgroep A+”

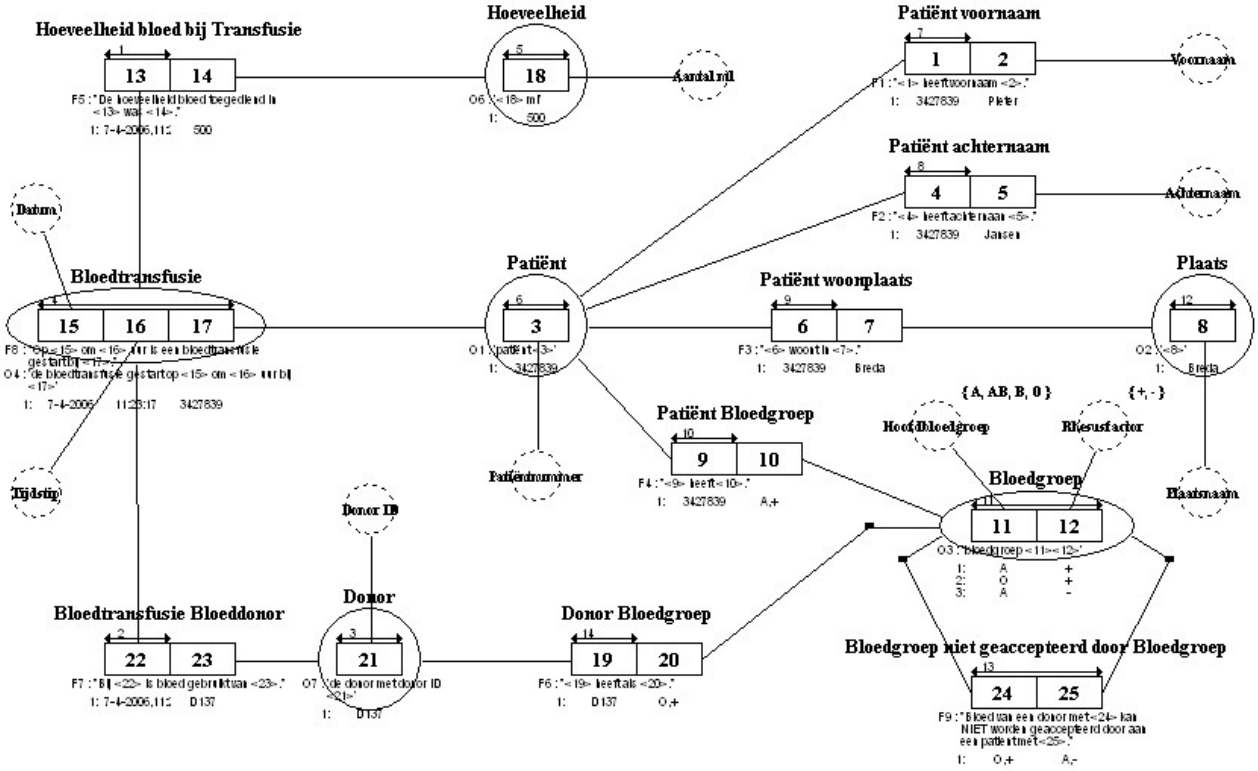
Evenzo wordt via een formulier dat gebruikt wordt bij een bloedtransfusie, de volgende gegevens vastgelegd:

Datum transfusie:	**7-4-2006**
Starttijd transfusie:	**11:23:17**
Hoeveelheid bloed (ml):	**500**
Donor	
Donor ID:	**D137**
Bloedgroep:	**O+**
Patiënt	
Patiëntnummer:	**3427839**
Bloedgroep:	**A+**

Hierbij zijn bijvoorbeeld als feiten te formuleren:

"Op 7-4-2006 om 11:23:17 uur is een bloedtransfusie gestart bij Patiënt 3427839"
"De hoeveelheid bloed toegediend in de bloedtransfusie gestart op 7-4-2006 om 11:23:17 uur bij Patiënt 3427839 was 500 ml"
"Bij de bloedtransfusie gestart op 7-4-2006 om 11:23:17 uur bij Patiënt 3427839 is bloed gebruikt van de donor met donor ID D137"
"De donor met donor ID D137 heeft als bloedgroep O+"
"Patiënt 3427839 heeft bloedgroep A+"

De laatste zin vonden we ook al bij het eerste formulier. Bij het gebruik van FCO-IM leiden deze zinnen tot de "elementaire informatie-grammatica" weergegeven in het diagram in figuur 2. Als we de grammatica bevolken met correcte concrete voorbeelden, dan kan het verkregen model alle zinnen over deze populatie terug genereren. Aan de hand van deze zinnen (feitformuleringen), kunnen domeindeskundigen - zoals artsen - het model volledig valideren. Deze methode wordt door ons bijvoorbeeld gevolgd bij het Erasmus Medisch Centrum in Rotterdam, waar wij een geïntegreerd Data Warehouse helpen bouwen voor de Intensive Care afdelingen. Dat een zorgvuldig gevalideerd model daarbij essentieel is, lijkt ons geen betoog te hoeven.

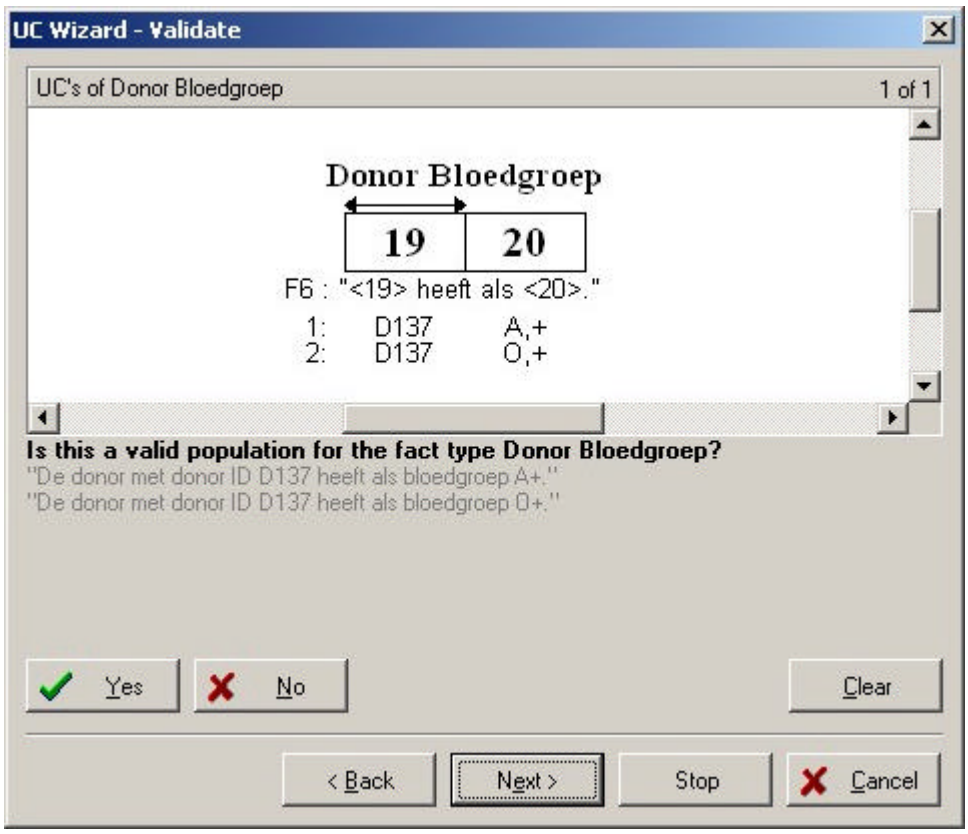


Figuur 2: ***De elementaire informatiegrammatica voor bloedgroepen***

Door het gebruik van FCO-IM verkrijgen we via natuurlijke taal in de eerste plaats een correcte tabelstructuur. Maar ook de beperkingsregels kunnen via communicatie met deskundigen bepaald worden. We doen dit door telkens concrete voorbeelden van feiten te geven en daarbij te vragen of die feiten mogelijk zijn. Zo kan bijvoorbeeld voor de bepaling van de 'uniciteitsconstraints', die uiteindelijk resulteren in primary of unique key constraints, een gestructureerd vraag- en antwoordspel gestart worden. In CaseTalk, de FCO-IM casetool waarmee FCO-IM geautomatiseerd ondersteund wordt [6] kan dit via een 'wizard', getoond in figuur 3. Daarin wordt gevraagd, of de gelijk-tijdige bevolking van het feittype 'Donor bloedgroep' geïmpliceerd door de twee feitformu-leringen

"De donor met donor ID 137 heeft als bloedgroep A+"
"De donor met donor ID 137 heeft als bloedgroep O+"

een geldige populatie is.



Figuur 3: ***De wizard voor het bepalen van uniciteitsconstraints***

De domeindeskundige heeft geen moeite met het beantwoorden van die vraag: “Nee, dat kan niet. Een mens heeft maar één bloedgroep.” CaseTalk vertaalt het antwoord ‘nee’ naar een uniciteits-constraint over rol 19: deze rol in het feittype ‘Donor bloedgroep’ kan maar één keer door dezelfde instantie van een donor ID bevolkt worden. In FCO-IM jargon heet dit: de rol in het feittype ‘Donor bloedgroep’ die gespeeld wordt door Donor (rol 19) heeft een enkelvoudige uniciteitsconstraint. Vervolgens zal de ‘wizard’ als vraag stellen, of de gelijktijdige bevolking van ‘Donor bloedgroep’ geïmpliceerd door de zinnen

“De donor met donor ID 137 heeft als bloedgroep A+”
“De donor met donor ID 921 heeft als bloedgroep A+”

een geldige populatie is. Hierop antwoordt de domeindeskundige nu natuurlijk bevestigend: “Ja, dat kan. Verschillende donors kunnen dezelfde bloedgroep hebben.” En CaseTalk plaatst dan geen uniciteitsconstraint over rol 20.

Daarmee is aan de noodzakelijke randvoorwaarde voor een valide datamodel voldaan, en wel door volledig gebruik te maken van communicatie. Maar met dit gebruik van communicatie bereiken we op zijn best een goede *data*kwaliteit, maar nog geen goede *informatie*kwaliteit en daar is het ons nu om te doen. Een valide datamodel kan weliswaar geen structureel incorrecte gegevens bevatten, maar nog wel *foutieve informatie*. Want informatiekwaliteit is een mensenzaak en mensen zijn in staat in een correcte structuur foutieve zaken op te slaan. Gelukkig kan communicatie ons ook helpen bij het realiseren van informatie-kwaliteit. Om dat aan te tonen nemen we een uitstapje naar het gebeuren in Japan.

Het informatiesysteem dat de verkoop van aandelen afhandelt, zou in vereenvoudigde vorm gebaseerd kunnen worden op een datamodel gegenereerd uit de volgende soort zinnen:

“De huidige koers van aandeel X547 is ¥ 500.000”
“Verkoop 327234 betreft de verkoop van aandeel X547”
“Verkoop 327234 betreft de verkoop van 10 stuk(s)”
“Verkoop 327234 gaat tegen een waarde van ¥ 510.000”

Uit deze zinnen kan een afleidbaar feit worden opgesteld:
“Verkoop 327234 levert een winst van ¥ 100.000”

Op het moment dat een gebruiker, zoals onze ongelukkige verkoper, zijn verkoop intypt, kan

het systeem zonder enige moeite de bijbehorende feiten terug genereren. Het is immers gebouwd op basis van die communicatie. Onze ongelukkige verkoper zou dan - *voordat de verkoop geëffectueerd wordt* - als feiten op zijn scherm hebben zien verschijnen:

"De huidige koers van aandeel J-com is ¥ 600.000"
"Verkoop 885694 betreft de verkoop van aandeel J-com"
"Verkoop 885694 betreft de verkoop van 610.000 stuk(s)"
"Verkoop 885694 gaat tegen een waarde van ¥ 1"
"Verkoop 885694 levert een winst van -/- ¥ 365.999.390.000"

Het komt ons voor, dat deze eenvoudige mededeling de onvoorstelbare ramp zou hebben voorkomen. Het is interessant om daarbij op te merken, dat er hiervoor niet allerlei dure slimheid in het programma hoeft te worden ingebouwd op grond waarvan het programma zelf deze verkoop zou moeten laten ketsen, zoals nu door de Mizuho Financial Group blijkbaar is gedaan. Dat kan de betrokken medewerker dan echt zelf wel af. En we laten aan het oordeel van de lezer over, welk type oplossing de eerbiedwaardige klant van het systeem werkelijk een volledig *mondige* klant maakt.

Hoe zit het nu met ons voorbeeld over bloedgroepen? Daar ligt de zaak wat gecompliceerder. Hoe voorkom je het gebruik van een donorhart of donorbloed met een verkeerde bloedtypering? Daarvoor moeten we ons oog voor informatiekwaliteit bij het modelleren wat verder aanscherpen. Nog onlangs rapporteerde CBS news een bericht over de nieuwste medische inzichten bij het transplanteren van harten. Dit bericht begint als volgt:

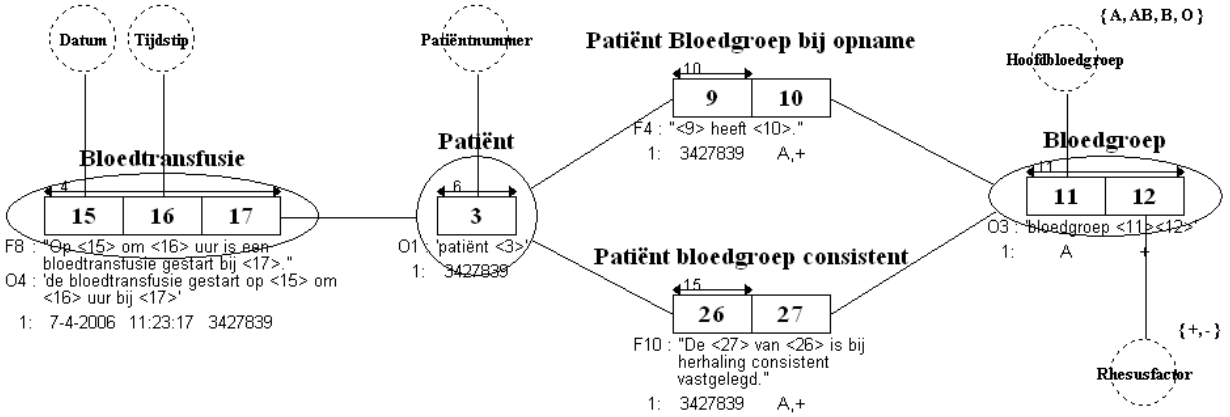
"Most any surgeon out of medical school is taught a rule that must never be violated: When transplanting a heart, the heart must be compatible with the patient's blood type. A mismatched heart leads to rapid death."

Nu gaan de nieuwste medische inzichten rond harttransplantaties ver voorbij de grens van onze medische kennis. Daarom richten we ons hier op bloedtransfusies waarin bovenstaand uitgangspunt als hard mag worden aangenomen. Succesvolle bloedtransfusies moeten simpelweg voldoen aan de bekende donor-acceptor tabel. Hiertoe dient het formulier bij de bloedtransfusie zowel voor de patiënt als de donor de juiste bloedgroep te bevatten. En deze bloedgroepen moeten compatibel zijn. Dat is letterlijk van levensbelang voor de patiënt. Daarom moeten we al bij het modelleren de informatiekwaliteit van deze gegevens voor ogen houden. Een standaard aanpak in dit soort situaties is om in elk geval af te dwingen, dat van iedere patiënt de bloedgroep is vastgelegd, bijvoorbeeld door de bloedgroep bij het opnameformulier verplicht te stellen, eventueel verkregen via een 'pick list' met geldige bloedgroepen. Datakwaliteit lijkt dan immers verzekerd: de kolom in de tabel is 'NOT NULL' en bevat altijd bruikbare gegevens. Dat is natuurlijk onzin. Hij is dan altijd gevuld, maar de *informatie*kwaliteit is daarmee nog verre van gegarandeerd: de informatie kan nog steeds onjuist zijn. Het is goed om hierbij te bedenken, dat gebruikers vaak een hekel hebben aan verplichte velden, zeker op het moment dat er druk op hen staat. Het zal duidelijk zijn, dat bij de Intensive Care afdeling in een ziekenhuis het redden van een mensenleven beslist een hogere prioriteit heeft dan het invullen van een verplicht veld. En uiteraard neemt men daar ook patiënten op waarvan de bloedgroep nog niet bekend is. Men wacht niet met behandeling tot dit bekend is, men vermijdt alleen handelingen die hiervan afhankelijk zijn, totdat men wel over de juiste informatie beschikt. In de tussentijd worden wel alle gegevens van de patiënt vastgelegd, waarvan men wel zeker is. Een 'NOT NULL' kolom werkt mogelijk het 'even' invullen van een tijdelijke waarde in de hand, met alle kans dat de daadwerkelijk correcte informatie nooit meer wordt ingevuld.

Simpelweg de kolom bloedgroep 'NOT NULL' maken helpt in dit geval de informatiekwaliteit dus in het geheel niet. Wat wel zal helpen, is de eis: als een patiënt een bloedtransfusie ondergaat dient de bloedgroep van de patiënt en de donor bekend en correct te zijn. Vastleggingssysteem voor bloedgroepen zullen daartoe doorgaans in de volgende trant zijn opgezet. Een bloedgroep van een patiënt of donor moet herhaald gemeten worden. Als een nieuw ingevoerde meting afwijkt van de vijf voorgaande metingen, geeft het systeem een waarschuwing af. Er moet dan opnieuw een onafhankelijke meting worden gedaan, totdat zeker is dat de nieuwe waarde consistent is. Pas dan wordt deze in het systeem vastgelegd. Op vergelijkbare manier kan ook het bloedtransfusiesysteem een signaal afgeven, als de twee bloedgroepen op het transfusieformulier niet compatibel zijn. Het is essentieel daarbij te bedenken, dat we die signalen steeds het beste

kunnen laten geven in de vorm van betekenisvolle feitformuleringen, en niet alleen in de vorm van rode signaallampjes of zo. Hiervoor is een iets subtielere vorm van het in figuur 2 getoonde data model noodzakelijk. De zin over de bloedgroep van de patiënt bij het opnameformulier mag blijven zoals hij is, maar de zin die hoort bij de bloedgroep van de patiënt op het transfusie-formulier luidt nu als volgt:

"De bloedgroep A+ van patiënt 3427839 is bij herhaling consistent vastgelegd."



Figuur 4: ***De uitbreiding van het model voor bloedtransfusies***

Dit leidt tot een uitbreiding van het model in figuur 2, zoals getoond in figuur 4. Als de dienstdoende staf bij de bloedtransfusie hun muis laten zweven over de bloedgroep van de patiënt in het transfusieformulier, dan krijgen ze bovenstaande geruststellende mededeling in plaats van de veel minder zeggende mededeling: "Patiënt 3427839 heeft bloedgroep A+". In FCO-IM termen betekent dit, dat er géén 'totaliteitsconstraint' (verplichte rol regel) wordt gezet op rol 9, maar wel een deelverzamelings-regel van rol 17 naar rol 26. Die regel zegt dan, dat elke patiënt die voorkomt in rol 17 in 'Bloedtransfusie' ook moet voorkomen in de verzameling patiënten in rol 26 van 'Patiënt bloedgroep consistent'. Oftewel, de patiënt-populatie van rol 17 in 'Bloedtransfusie' is een deelverzameling van de patiëntpopulatie van rol 26 in 'Patiënt bloedgroep consistent'.

Als dan de bloedgroepen van de patiënt en de donor beide bekend zijn, kan het systeem simpelweg checken of de combinatie van deze bloedgroepen niet voorkomt in het feittype 'Bloedgroep niet geaccepteerd door Bloedgroep' in figuur 2. Doet hij dat wel, zoals het geval zou zijn, als bij onze patiënt Pieter Jansen een bloedgroep A- zou zijn aangegeven, dan geeft het systeem de mededeling:

"De donor bloedgroep O+ kan NIET worden geaccepteerd door een patiënt met bloedgroep A-."

Dit is precies de verwoording die is uitgesproken bij de analyse van de donor-acceptor tabel. Mocht de dienstdoende staf nog willen weten *waarom* deze melding verschijnt, dan worden zij weer geholpen door zinnen die bij de modellering zijn uitgesproken:

"Bij de bloedtransfusie gestart op 7-4-2006 om 11:23:17 uur bij Patiënt 3427839 is bloed gebruikt van de donor met donor ID D137"
"De donor met donor ID D137 heeft als bloedgroep O+"
"Patiënt 3427839 heeft bloedgroep A-"

En de dienstdoende staf zal overeenkomstig reageren en om een andere zak bloed vragen.

Met de hier voorgestelde aanpak hebben we op een moderne manier invulling gegeven aan de ideeën van 'Total Information Quality Management' van Larry English. Eén van zijn belangrijkste principes is, dat het aanbrengen van *data*kwaliteit in gegevensverzamelingen op zich niet genoeg is, maar pas effect sorteert, als daardoor het kwaliteitsgedrag van de gebruikers in positieve zin wordt beïnvloed. En we hebben zo ook recht gedaan aan de uitspraken van Wittgenstein [2, 8]): *"Die Welt ist die Gesamtheit der Tatsachen, nicht der Dinge..."*:

Het bestaan van een steen in een woestijn waar nooit een mens komt, heeft voor geen enkel mens enige zinnige betekenis. Wil het bestaan van de steen betekenis krijgen, dan moet er iemand geweest zijn om dat bestaan vast te stellen en te communiceren aan anderen. Pas dan heeft zelfs een ezel er wat aan. In de geest van Wittgenstein is dan ook de steen *als ding* zelf niet relevant, maar wel *het feit*, dat het bestaan van die steen

communiceert aan gebruikers van de woestijn. Evenzo is informatiekwaliteit die men in een gegevensverzameling stopt, zonder betekenis, als men deze niet openlijk communiceert naar gebruikers van die informatie en daarmee tastbare en bruikbare feiten produceert, die het juiste kwaliteitsgedrag bij de vastlegging van gegevens en het gebruik van de informatie stimuleren.

Referenties

1. Database Magazine, jaargang 16, no. 7 (december 2005).
2. Database Magazine, jaargang 16, no. 2 (maart 2005).
3. Database Magazine, jaargang 11, no. 8 (december 2000).
4. Database Magazine, jaargang 12, no. 1 (januari 2001).
5. Volledig Communicatiegeoriënteerde Informatiemodellering, G. Bakema, J.P. Zwart, H. van der Lek, Kluwer BedrijfsInformatie, 1996
6. CaseTalk 6.5, Bommeljé, Crompvoets en Partners, 2005
7. Improving Data Warehouse and Business Information Quality, Larry P. English, John Wiley & Sons Inc., 1999
8. Tractatus logico-philosophicus, Ludwig Wittgenstein, Atheneum –Polak & Van Gennep, 1976

Over de Auteurs

Dr. Peter W.F. Alons is senior consultant bij Atos Origin/BI-CRM en ruim vijftien jaar betrokken geweest bij een groot aantal Business Intelligence en Data Warehouse projecten bij diverse bedrijven.
Drs. Ing. Rob G. Arntz is eveneens consultant bij Atos Origin/BI-CRM en ruim vijf jaar betrokken bij diverse Business Intelligence en Data Warehouse projecten.

Contact

Peter Alons: Peter.Alons@AtosOrigin.Com
Rob Arntz: Rob.Arntz@AtosOrigin.Com
Atos Origin: www.atosorigin.com