Gegevenskwaliteit in overheidsnetwerken

Carolien Besselink

SAMENVATTING Het delen en uitwisselen van gegevens is nodig om de dienstverlening door de overheid te kunnen verbeteren. Maar dat werkt alleen als de gegevens goed zijn. Bij uitvoeringsorganisaties bepaalt de kwaliteit van de ontvangen gegevens grotendeels de efficiëntie van het verwerkingsproces. Niet correcte gegevens genereren foute beschikkingen voor burgers en bedrijven. Dat schaadt het imago van de overheid.

Het belang van gegevenskwaliteit voor het functioneren van ketens en netwerken wordt in veel Nederlandse beleidsstukken genoemd, maar niet hoe de eigenaar van een gegeven daar zorg voor draagt. In de praktijk zijn er veel problemen bij het uitwisselen van gegevens. Om een stap verder te komen met deze problematiek is het nodig dat we gegevens als een zelfstandig product gaan zien en beschrijven. Zoals kleur, topsnelheid of degelijkheid een auto beschrijven, zo kan ook een bericht beschreven worden met kenmerken als syntactische juistheid, variabiliteit of begrijpelijkheid.

RELEVANTIE VOOR DE PRAKTIJK Dit artikel reikt een model aan dat managers, adviseurs en auditors handvatten biedt voor het aanpakken van problemen bij het uitwisselen van gegevens in overheidsnetwerken. Het gaat in op de factoren die de kwaliteit van gegevens bedreigen en te treffen maatregelen. Daarbij helpen specifiek voor gegevens gedefinieerde kwaliteitskenmerken. De praktijkcasus van de loonaangifteketen laat zien hoe belangrijk het is om gegevenskwaliteit goed te kunnen omschrijven.

1 Inleiding

De laatste tien jaar heeft de Nederlandse overheid tal van programma's gelanceerd om de kwaliteit van de dienstverlening te verhogen door middel van digitalisering. De elektronische overheid ontwikkelt zich richting een netwerk van overheden die continu met elkaar in verbinding staan om gegevens uit te wisselen. Naarmate het aantal gebruikers van een gegevensverzameling toeneemt, wordt het belangrijker dat deze verzameling een hoge kwaliteit heeft. Het werken in netwerken stelt hogere eisen aan uit te wisselen gegevens dan het werken in ketens. Bij netwerken is strakke sturing minder goed mogelijk dan bij ketens, want de band tussen partners is minder sterk en de uitwisseling is in mindere mate vooraf bepaald. Als vraag en aanbod van gegevens frequent wisselen, is het maken van procedureafspraken moeilijker. Daarom komt het bij netwerken, meer nog dan bij ketens, aan op de kwaliteit van de berichten zelf. Een bericht moet op zichzelf staan, zelfverklarend en zelfstandig bruikbaar zijn. In een netwerk moet de gebruiker de kwaliteit van een bericht vast kunnen stellen aan de hand van de inhoud en de kenmerken van het bericht zelf.

Door deze beweging naar een 'connected world' staat de kwaliteit van gegevens weer volop in de belangstelling. In de Verenigde Staten is de laatste tien jaar veel onderzoek gedaan en zijn onder andere door het Massachusetts Institute of Technology *data quality*-programma's gestart. De Europese Unie geeft met de *Directive on the Re-use of Public Sector Information* (2003) het belang van hergebruik van gegevens aan. Het *European Interoperability Framework* (2004) gaat uit van het subsidiariteitsbeginsel. Daarmee legt de Europese Commissie de nadruk op het maken van afspraken die betrekking hebben op de koppelvlakken tussen organisaties. Het Forum Standaardisatie stelt het uit kunnen wisselen van gegevens centraal in de Interoperabiliteitsagenda (2008) en in een bundel met artikelen (Zwienink en Wisse, 2008). De ICT Uitvoeringsorganisatie geeft in de Nederlandse Overheid Referentie Architectuur 2.0 (NORA, 2007) aan dat elk gegeven een eigenaar moet hebben en dat deze eigenaar verantwoordelijk is voor de kwaliteit. Als er gegevens geleverd worden, dan moet de kwaliteit ervan bekend zijn. De wetgever heeft met de basisregistraties vastgelegd wat 'authentieke' gegevens zijn. Deze basisgegevens worden eenmaal uitgevraagd bij burgers en bedrijven en andere overheden worden verplicht om ze te gebruiken.

De gevolgen van deze ontwikkelingen zijn ingrijpend voor organisaties gericht op massale gegevensverwerking, zoals de Belastingdienst en het UWV. Starreveld (2006) kenmerkt

MAB04_binnen.indd 209 29/03/10 4:51 PM

deze organisaties met het aanwezig zijn van massaal repeterende processen die resulteren in klantspecifieke informatieproducten. Zowel de invoer, de tussenproducten als de eindproducten zijn gegevens. Voorbeelden van producten uit deze informatiefabrieken zijn een belastingaanslag of een uitkeringstoezegging. Het uitwisselen van gegevens in netwerken geeft heel nieuwe mogelijkheden voor deze organisaties, bijvoorbeeld voor toezicht. Maar tevens ontstaat er een grote onderlinge afhankelijkheid. Want vervuilde gegevens bij één van de organisaties kunnen vergaande gevolgen hebben voor meerdere afnemers in het netwerk.
Voor het investeren in gegevenskwaliteit is een businesscase te maken, waarbij de kosten voor het verbeteren van de kwaliteit afgezet worden tegen de baten. Er zijn kosten voor preventie en kosten voor het corrigeren van foute gegevens. De baten zijn een vermindering van deze arbeidskosten en een toename van de inkomsten. Door bijvoorbeeld de gegevens van verschillende instanties te combineren, kan belastingfraude opgespoord worden. Daarnaast kan een goede kwaliteit van de ontvangen gegevens de processen verkorten en de productiviteit verhogen.

Gegevenskwaliteit is als onderzoeksgebied nog volop in ontwikkeling. Oude begrippen als juistheid, tijdigheid en volledigheid blijken niet meer afdoende om de hedendaagse problematiek te adresseren. Dit artikel is gebaseerd op een onderzoek (Besselink, 2009) naar de factoren die de kwaliteit van gegevens beïnvloeden bij het uitwisselen in overheidsnetwerken. In dit onderzoek wordt eerst nagegaan wat de literatuur zegt over gegevenskwaliteit en met welke kenmerken deze beschreven kan worden. Daarna is onderzocht wat de kwaliteit van gegevens bedreigt en welke maatregelen mogelijk zijn. De definities van ook nieuwe kwaliteitskenmerken, de bedreigingen en de mogelijke verbeteracties zijn vervolgens geijkt aan de praktijksituatie van de loonaangifteketen. Het praktijkonderzoek vond plaats door interviews en bestudering van documenten. Het toepassen van de resultaten uit de literatuurstudie in de praktijk bevestigde het nut van een begrippenkader om de kwaliteit van gegevens aan te duiden.
Het onderzoek beperkte zich tot het uitwisselen van gegevens binnen de overheid, want daar gelden richtlijnen voor informatiearchitectuur en is uitwisseling veelal gebaseerd op wetgeving. Maar de in dit artikel gepresenteerde aanpak is evenzeer bruikbaar voor niet-overheidsbedrijven. Niet in beschouwing genomen zijn privacy- en beveiligingsaspecten, en evenmin de persoonlijke of financiële belangen die gemoeid zijn met de registratie van gegevens.
Dit artikel beschrijft het tijdens het onderzoek ontwikkelde model de 'Datadriehoek'. Dit model onderscheidt drie invalshoeken voor het beheersen van de kwaliteit van gegevens. Paragraaf 2 beschrijft de belangrijkste kwaliteitskenmerken, past deze toe op de praktijk van de loonaangifte en laat een voorbeeld zien van het afstemmen van vraag en aanbod van gegevens. Paragraaf 3 gaat in op wat gegevenskwaliteit bedreigt en paragraaf 4 op wat de verbetermogelijkheden zijn. Paragraaf 5 beschrijft de Datadriehoek als geheel en paragraaf 6 ten slotte bevat de conclusies.

2 Gegevenskwaliteit

2.1 Definities en kwaliteitskenmerken

Het verbeteren van de kwaliteit van gegevens begint bij het begrijpen wat goede gegevens zijn, bij het definiëren van wat de kwaliteit van een informatieproduct is. Onder een informatieproduct wordt een set digitale gegevens in een 'vast te pakken vorm' verstaan, zoals een XBRL-bericht. De ISO 8000-norm voor data quality is nog in ontwikkeling, dit artikel hanteert daarom de definities van McGilvray (2008):
Gegevens (data): herinterpreteerbare weergave van informatie op een geformaliseerde manier geschikt voor communicatie, interpretatie of verwerking.
Gegevenskwaliteit: de mate waarin gegevens een betrouwbare bron vormen voor één of meerdere toepassingen.

De eigenaar van een gegeven is verantwoordelijk voor de kwaliteit van de gegevens in zijn verzameling. Hij bepaalt wat er vastgelegd wordt en of een wijziging doorgevoerd mag worden. Ook geeft hij aan wat de mogelijke afwijkingen in zijn gegevens zijn. Bij een auto bepaalt de fabrikant de betrouwbaarheid en het imago van het product. Ook bij informatieproducten bepaalt de leverancier het kwaliteitsniveau. Zoals de ene persoon een snelle auto wil en de ander meer hecht aan een veilige auto, zo heeft de ene organisatie heel nauwkeurige gegevens nodig en een andere vooral volledige. Om een idee te krijgen van de geschiktheid van een database of bericht moeten we dus de kenmerken aanduiden die voor een toepassing nodig zijn. We moeten deze kwaliteitskenmerken gaan benoemen, beschrijven en meten.
Maar voor gegevens zijn er nog geen algemeen geaccepteerde definities, er zijn honderden kwaliteitskenmerken te definiëren (Redman, 2004). Onderzoekers zijn het zelfs niet eens over welke kenmerken onderscheiden zouden moeten worden. ISO (2008) werkt aan normen welke ingaan op *accuracy*, *completeness* en *provenance*. De NORA (2007) noemt als kwaliteitskenmerken actualiteit, accuraatheid en betrouwbaarheid. De *Directive on the Re-use of Public Sector Information* (2003) gaat in op het belang van actuele gegevens die tijdig beschikbaar gesteld worden. Geen van deze publicaties geeft definities. De beroepsorganisatie van IT-auditors NOREA (1998) heeft zeven algemene kwali-

MAB04_binnen.indd 210 29/03/10 4:51 PM

Tabel 1 Kwaliteitskenmerken voor gegevensuitwisseling in overheidsnetwerken

Kenmerk	Definitie
Juistheid	
Syntactisch	**De mate van overeenstemming van het formaat en/of de waarde van het veld met hetgeen toegestaan is (naar Batini en Scannapieco, 2006 en McGilvray, 2008).** Controle of een datumveld inderdaad een datum bevat en een numeriek veld geen letters. Onder syntax valt ook de controle of het veld één van de toegestane waarden in het domein bevat. De syntactische juistheid wordt beoordeeld met vergelijkingsfuncties die de afstand evalueren tussen de waarde in de database en de waarden in het domein. Meten: tellen van het aantal *deletions, insertions* en *replacements* dat de waarde in het veld omzet in een toegestane waarde. Of: het aantal syntactisch juiste velden delen door het totaal aantal velden (%).
Semantisch	**De mate waarin een gegeven overeenkomt met de afgebeelde werkelijkheid (Batini en Scannapieco, 2006).** Om de semantische nauwkeurigheid te kunnen beoordelen moet de werkelijke waarde bekend zijn of uit additionele kennis (bijvoorbeeld uit brondocumenten) afgeleid kunnen worden. De semantische nauwkeurigheid wordt uitgedrukt met <correct, niet correct>. Meten: het aantal correcte velden delen door het totale aantal velden (%).
Bedrijfsregels	**De mate waarin voldaan wordt aan regels die volgen uit het beleid (Batini en Scannapieco, 2006).** Deze regels zijn context afhankelijk. Bijvoorbeeld: 'als de burgerlijke staat gehuwd is, dan mag de leeftijd niet lager dan 18 jaar zijn'. Meten: het aantal gecontroleerde gevallen waarin de bedrijfsregels gerespecteerd werd delen door het totaal aantal gecontroleerde gevallen (%).
Integriteitsregels	**Een maat voor het bestaan, de geldigheid, de structuur, de inhoud en andere basis karakteristieken van gegevens (McGilvray, 2008).** Bij relationele databases gaan *intrarelation constraints* over attributen binnen één tabel en *interrelation constraints* over attributen in verschillende tabellen. Meten: het aantal gecontroleerde gevallen waarin de integriteitregels gerespecteerd werd delen door het totaal aantal gecontroleerde gevallen (%).
Volledigheid	
Dekkingsgraad	**De mate waarin een database waarden bevat over de populatie waarin men geïnteresseerd is (McGilvray, 2008 en Batini en Scannapieco, 2006).** Meten: het aantal weergegeven objecten in de gegevensverzameling delen door het totaal aantal objecten in de populatie (%).
Vullingsgraad	**De mate waarin alle velden in een regel, kolom of tabel gevuld zijn met een waarde (Batini en Scannapieco, 2006).** Meten: het aantal gevulde velden (niet-nul waarden) delen door het totaal aantal velden (%).
Tijdgerelateerd	
Actualiteit	**De mate waarin gegevens prompt worden aangepast (Batini en Scannapieco, 2006). Vooral van belang voor gegevens die snel verouderen, bijvoorbeeld een bestand met adressen.** Meten: het tijdsinterval van binnenkomst van een wijziging tot het moment dat deze wijziging opgenomen is in de database.
Tijdigheid	**De mate waarin tijdig over gegevens beschikt kan worden (Batini en Scannapieco, 2006).** Meten: er zijn complexe formules voor (Batini en Scannapieco, 2006), maar voor sommige toepassingen kan de ouderdom van de gegevens een maat zijn.
Bron	
Toegankelijk	**De mogelijkheid voor een gebruiker om de bron te benaderen (Pipino, 2006).** Meten: Pipino (2006) geeft een complexe formule met drie parameters: 1) het tijdsinterval van vraag tot beschikbaarheid van de gegevens, 2) het tijdsinterval totdat beschikbaar stellen van de gegevens geen zin meer heeft en 3) een gevoeligheidsparameter.
Vertrouwenswaardig	**De mate waarin een bron betrouwbaar wordt geacht. Wordt wel opgebouwd uit geloofwaardigheid, reputatie en objectiviteit (Pipino, 2006).** Recentelijk wordt in *peer-to-peer* communicatie een 'trust level' bepleit, dat verbonden wordt aan elke peer voor ieder type gegevens dat het levert.
Schema	**Kwaliteit van het schema: leesbaar, volledig, correct en zich beperken tot de essentie (Batini en Scannapieco, 2006).**
Bericht	
Begrijpelijk	**De mate van beschikbaarheid van documentatie en metadata om het bericht correct te kunnen interpreteren (naar Batini en Scannapieco, 2006; Pipino, 2006; en McGilvray, 2008).**
Waarachtig	**De mate waarin het bericht als waar en geloofwaardig beschouwd wordt (Pipino, 2006).** Dit is een subjectief criterium en te meten via vragenlijsten. Meten: Pipino (2006) drukt de waarachtigheid uit als de minimumwaarde van een van de volgende drie waarden: 1) geloofwaardigheid zoals gemeten met vragenlijsten, 2) standaardwaarde/ algemeen geaccepteerde waarde voor waarachtigheid en 3) geloofwaardigheid gebaseerd op de ouderdom van de gegevens.

MAB04_binnen.indd 211 29/03/10 4:51 PM

Kenmerk	Definitie
Relevant	De waarde van gegevens en de mate waarin deze van toepassing zijn voor een specifiek proces van een afnemer (naar McGilvray, 2008).
Verwerkbaar	De mate waarin de gegevens tot de gewenste bedrijfstransacties of uitvoer zullen leiden (McGilvray, 2008).
Variabiliteit	De mate waarin en de frequentie waarmee gegevens in de tijd veranderen (naar Batini en Scannapieco, 2006 en McGilvray, 2008).
Synchroon	De mate van overeenstemming tussen gebruikte of opgeslagen gegevens in verschillende gegevensverzamelingen (McGilvray, 2008).

teitskenmerken gedefinieerd, maar voor het beschrijven en meten van de kwaliteit van gegevens bieden deze kenmerken te weinig houvast.
Om toch verder te komen met de kwaliteit van de gegevens die tussen overheden uitgewisseld worden, doet dit artikel een voorstel voor een beperkte set van kwaliteitskenmerken. Deze kenmerken zijn beschreven in tabel 1 en zijn gebaseerd op literatuur over gegevenskwaliteit (Batini en Scannapieco, 2006; McGilvray, 2008; Pipino et al., 2006). Gezamenlijk bieden deze kenmerken een begrippenkader om over de kwaliteit van gegevens te kunnen communiceren.

De tabel begint met de bekende kwaliteitskenmerken juistheid, volledigheid en de tijdgerelateerde criteria. Het is zinvol deze te verbijzonderen. De juistheid van gegevens wordt bepaald door de syntactische en semantische juistheid en door het voldoen aan bedrijfs- en integriteitregels. Volledigheid is de resultante van de dekkingsgraad en de mate waarin alle velden van een gegevensverzameling gevuld zijn. Voor afnemers is de tijdige beschikbaarheid van actuele gegevens belangrijk.
Het kwaliteitskenmerk juistheid adresseert onder andere het probleem van het dubbel voorkomen van objecten in een informatieproduct. Bij volledigheid speelt onder andere de vraag of een 'o' in een veld betekent dat de waarde o is of dat er geen waarde bekend is. Tijdgerelateerde kenmerken zijn van belang omdat soms niet gewacht kan worden met het starten van een proces. Zijn bijvoorbeeld de saldigegevens van banken tijdig beschikbaar om opgenomen te worden in de vooringevulde aangifte inkomstenbelasting?
Recent onderzoek richt zich met name op kwaliteitskenmerken die bij gegevensuitwisseling in netwerken belangrijk zijn. Deze nieuwe aanvullende kenmerken gaan over de bron, het schema en het bericht zelf. Betreffende de bron speelt het toegankelijk zijn van de gegevens en het kunnen vertrouwen op de gegevensleveringen vanuit de desbetreffende organisatie. Het schema van een informatieproduct moet goed leesbaar zijn, volledig, correct en zich beperken tot de essentie. Een afnemer hecht waarde aan een begrijpelijk, waarachtig en relevant bericht dat zijn processen kunnen verwerken. De mate waarin gegevens veranderen, is een kenmerk om rekening mee te houden, vooral als meerdere gegevensverzamelingen synchroon gehouden moeten worden.
Een voorbeeld van het belang van deze nieuwe kwaliteitskenmerken is de vertrouwenswaardigheid van de bron. In de Gemeentelijke Basis Administratie van Amsterdam blijkt 14,1 procent van de Amsterdammers op een ander adres te wonen dan geregistreerd is (Automatiseringsgids, 24 april 2009). Adresgegevens veranderen snel. Stabiele gegevens veranderen niet en hebben een variabiliteit van 0, bijvoorbeeld een geboortedatum.

2.2 Kwaliteitskenmerken in de loonaangifteketen

De situatie in de loonaangifteketen laat zien dat de praktijk een begrippenkader voor gegevenskwaliteit nodig heeft. Het doel van het verzamelen van gegevens in de loonaangifteketen is het kunnen leveren van gegevens aan de afnemers. De verzamelde gegevens worden bijvoorbeeld gebruikt bij het vaststellen van uitkeringen en toeslagen. De hoofdrolspelers in deze keten zijn het Uitvoeringsinstituut Werknemersverzekeringen (UWV), een zelfstandige bestuursorganisatie onder verantwoordelijkheid van de minister van SZW, en de Belastingdienst (BD), die ressorteert onder de minister van Financiën. Sinds een wetswijziging van 2006 is het takenpakket van de BD uitgebreid met de heffing en inning van de premies werknemersverzekeringen (WW, WAO enzovoort). De BD ontvangt 'collectieve' gegevens van inhoudingsplichtigen zoals werkgevers en uitkeringsinstanties. Het UWV beheert de Polisadministratie (Polis), de database met inkomensgegevens op naam ('nominatief'). Een nominatieve loonaangifteregel bevat per burgerservicenummer circa 120 gegevenselementen.
Met de wetswijziging is een wederzijdse afhankelijkheid ontstaan tussen UWV en BD, waarvan de breedheid en complexiteit onderschat zijn. Er waren flinke problemen die aan de Tweede Kamer uitgelegd moesten worden. Sindsdien heeft gegevenskwaliteit de volle aandacht van de bewindslieden: 'Om de werkende keten te realiseren, is de kwaliteit van de gegevens een belangrijke voorwaarde', aldus de minister van SZW en de staatssecretaris van Financiën in hun brief aan de Tweede Kamer van april 2008.

MAB04_binnen.indd 212 29/03/10 4:51 PM

Een integrale analyse van de problemen in de loonaangifteketen (IPA, 2007) noemt volledigheid, juistheid en tijdigheid als de belangrijkste criteria om de keten te beheersen. Daarbij wordt gewerkt met de volgende omschrijvingen: 'Juistheid: het borgen van de kwaliteit door middel van kwaliteitssignalen van de afnemers en het toezichtproces van de Belastingdienst. Volledigheid: alle afgesproken gegevens over de inschrijving van de inhoudingsplichtige en alle door de Belastingdienst van de inhoudingsplichtige ontvangen werknemersgegevens, conform de gegevensset van de loonaangifte, zijn beschikbaar voor afname in de Polisadministratie. Tijdigheid: de werknemersgegevens zijn op de zesde werkdag na ontvangst van de loonaangifte door de Belastingdienst beschikbaar voor afname uit de Polisadministratie'.

Bij deze omschrijvingen van juistheid, volledigheid en tijdigheid lopen definities, normen en processtappen door elkaar. De omschrijving van juistheid beschrijft de aanpak, niet wat juistheid is. De definitie van volledigheid raakt niet de dekkingsgraad. Mogelijk zijn alle ontvangen werknemersgegevens wel voor afname beschikbaar, maar als slechts een klein percentage inhoudingsplichtigen aangifte gedaan heeft, zullen afnemers de levering niet als volledig ervaren. De gegeven omschrijving van tijdigheid is een norm voor de doorlooptijd van een intern proces, geen definitie.

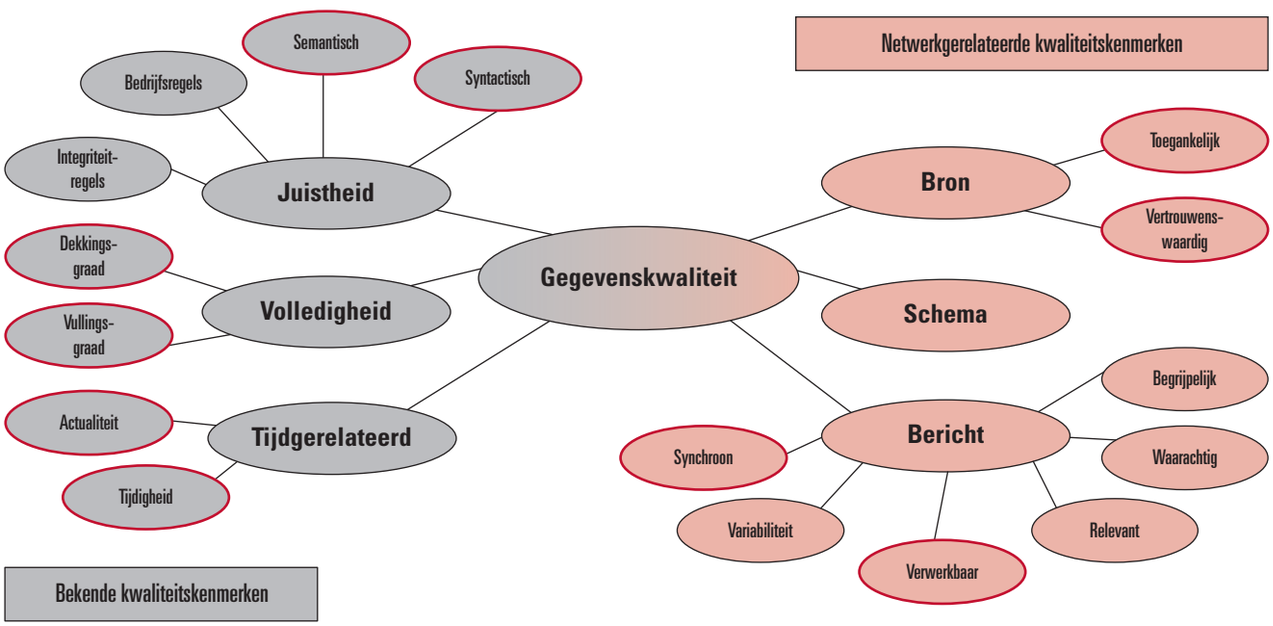
De loonaangifteketen worstelt met gegevenskwaliteit. Door de grote hoeveelheidgegevens en het niet tijdig aanleveren door inhoudingsplichtigen is het volgens de ketenmanager niet mogelijk om garanties te geven over de kwaliteit. 'De consequentie daarvan is dat het een fictie is te veronderstellen dat alle afnemende processen altijd kunnen beschikken over alle gegevens die ze nodig hebben' (Sturing op de kwaliteit van gegevens, 2008). 'Daarnaast ontbreekt een toetssteen voor wat de te behalen kwaliteit zou moeten zijn. (.....) Een waarborg voor een zo goed mogelijke kwaliteit ...' (Normenkader werkende loonaangifteketen, 2008).

Wie koopt een auto op het woord van de leverancier dat de kwaliteit zo goed mogelijk is? Zonder specificaties of garanties? Van de andere kant, de loonaangifteketen handelt naar het niveau van de kennis in Nederland. Er zijn op dit moment geen standaarden voor het beschrijven van gegevenskwaliteit en van een praktijksituatie mag niet het ontwikkelen van een theoretisch kader hiervoor verwacht worden.

Kan met de verbijzondering van kwaliteitskenmerken uit paragraaf 2.1 meer exact omschreven worden wat er aan de hand is met de kwaliteit van gegevens in de loonaangifteketen? Om zicht te krijgen op de kwaliteitskenmerken die spelen, is *reversed engineering* toegepast. Uit de problemen en de getroffen maatregelen in de loonaangifteketen zijn op basis van de definities in tabel 1 de kwaliteitskenmerken geïdentificeerd. In de keten ging de aandacht tot eind 2008 vooral uit naar de kenmerken die met een geaccentueerd kader zijn weergegeven in figuur 1.

Het bereiken van syntactisch juiste gegevensverzamelingen bleek niet eenvoudig in de loonaangifteketen. Nummers moeten omgezet worden en vertaaltabellen zijn niet

Figuur 1 Belangrijke kwaliteitskenmerken voor gegevens in de loonaangifteketen tot eind 2008



MAB04_binnen.indd 213 29/03/10 4:51 PM

helemaal sluitend. Het begrip inhoudingsplichtige van de BD komt niet geheel overeen met het begrip werkgever dat het UWV hanteert. De keten beschouwt de aangeleverde gegevens als juist als aan de syntactische eisen voldaan wordt. Het toezichtproces van de Belastingdienst moet de semantische juistheid waarborgen. Qua volledigheid kampt de loonaangifteketen met zowel de dekkingsgraad als de vullingsgraad. Het streven van de keten is het ontvangen van aangiften van álle inhoudingsplichtigen, dat wil zeggen 100 procent dekkingsgraad. Hierbij spelen de vragen of de gehele populatie van inhoudingsplichtigen een uitnodiging tot het doen van aangifte ontvangen heeft, welk deel van deze populatie daadwerkelijk aangifte doet en welk deel hiervan door de controles komt en opgenomen wordt in Polis. Het incompleet aanleveren van gegevens en de uitval tijdens het verwerkingsproces leiden tot een onvolledige vulling van Polis. Van de tijdgerelateerde kenmerken heeft het prompt verwerken van aan- en afmeldingen, het tijdig toekennen van loonheffingnummers en het tijdig in- en uitschrijven van werkgevers invloed op de actualiteit.
Bij de netwerkgerelateerde kenmerken speelt toegankelijkheid tot de gegevens onder andere bij het realiseren van een inkijkfunctie in Polis. Ook de vertrouwenswaardigheid van de bron is een probleem: 'Polis is nog niet stabiel, robuust en onderhoudbaar, er is geen gedegen beheer' (Smits, 2008). Door de vele koppelvlakken en verwerkingsslagen is de mate van overeenstemming van gegevens in de verschillende databases een ware uitdaging voor de keten. De verwerkbaarheid van gegevensleveringen uit Polis is bij veel afnemers in onderzoek.

De loonaangifteketen komt niet zonder slag of stoot tot stand, maar er worden belangrijke doelen mee bereikt. Zo zijn de administratieve lasten van werkgevers verminderd doordat er maar één loket meer is en werkgevers minder enquêtes van het CBS hoeven in te vullen. De uitvoeringskosten zijn lager en de loonheffing en het toezicht erop lopen goed. Dat leidt tot hogere belastinginkomsten.
Van het belang van goede gegevens is de loonaangifteketen doordrongen, echter het ontbreken van een eenduidig begrippenkader vertroebelt de discussie. Vanaf 2010 zullen ook andere afnemers betrokken worden bij de kwaliteit van de gegevens. Zij zullen berichten willen begrijpen, beoordelen op relevantie en waarachtigheid en willen weten wat de frequentie van wijziging is. Dan spelen alle kwaliteitskenmerken in figuur 1 een rol.

2.3 Kwaliteitsprofielen

Gegevens zijn goed voor afnemers als de verwerking ervan probleemloos tot goede tussen- en eindproducten leidt. Aan welke eisen moeten de gegevens dan voldoen? Een fysiek product als een bout wordt beschreven door productspecificaties als het materiaal, de lengte van de schroefdraad en de diameter. Voor de diameter geldt een nominale waarde en afhankelijk van de kwaliteit een afwijking (tolerantie) met een minimum- en maximumwaarde. English (2002) benadrukt dat we ook bij informatieproducten met specificaties moeten werken die zowel de inhoud als de kwaliteitskenmerken weergeven.
De specificaties van een bericht of database kunnen vastgelegd worden met een kwaliteitsprofiel. Het *kwaliteitsprofiel* van een informatieproduct specificeert de gegevenselementen die het bevat, de kwaliteitskenmerken per gegevenselement, de streefwaarde per kenmerk en de tolerantie die daarbij wordt toegestaan. Bij het uitwisselen van gegevens gaat het dan om de vraag of het kwaliteitsprofiel van de aangeboden gegevens past binnen de toleranties van het kwaliteitsprofiel van de gevraagde gegevens.
Om het afstemmen van vraag en aanbod van gegevens te verduidelijken, is figuur 2 opgenomen. Dit voorbeeld is afgeleid uit de praktijk van de loonaangifteketen, maar is vereenvoudigd en deels fictief. Het beperkt zich tot slechts enkele gegevenselementen met enkele kwaliteitskenmerken en voor de eenvoud zijn bovendien de toleranties buiten beschouwing gelaten. Bij dit voorbeeld gelden de definities uit tabel 1.

2.3.1 Aanbod

Het voorbeeld beschrijft een potentiële overdracht van gegevens van Organisatie 1 naar de Organisaties 2, 3 en 4. De linkerzijde van de figuur geeft een beeld van de kwaliteit van de leverancier van de gegevens, zijnde Organisatie 1, en van de gegevens zelf. De kwaliteitskenmerken 'toegankelijk' en 'vertrouwenswaardig' kenschetsen de bron. In het voorbeeld is de toegankelijkheid beperkt, want de gegevens mogen alleen worden verstrekt als daarvoor een wettelijke basis is. De mate waarin afnemers in het algemeen de kwaliteit van gegevens van Organisatie 1 beoordelen, is gemeten met een enquête onder afnemers. Dat leverde een score op van 0,80 op een schaal van 0 tot 1.
Organisatie 1 heeft mogelijk meerdere gegevensverzamelingen, maar het voorbeeld gaat over de Database Inkomensgegevens. De afnemers vinden de inhoud van deze database heel begrijpelijk, zij beoordelen het met een score van 0,95. Mogelijk is er een helder en goed gedocumenteerd datamodel beschikbaar. De waarachtigheid waarderen de afnemers iets lager met 0,80.
De Database Inkomensgegevens bevat vijf gegevenselementen: de Betrokkene, de Werkgever of uitkeringsinstantie, de Sectorcode, het Soort inkomen en het Bedrag cumulatief inkomen. Het cumulatief inkomen is het inkomen van een betrokkene in een bepaald jaar en wordt opgebouwd door het jaar heen. In januari is het Bedrag cumulatief inkomen slechts één maandinkomen, terwijl het eind december de som kan zijn van twaalf salarisuitbetalingen, vermeerderd met het uitgekeerde vakantiegeld.

MAB04_binnen.indd 214 29/03/10 4:51 PM

Figuur 2 Voorbeeld van het afstemmen van kwaliteitsprofielen

KWALITEITSPROFIELEN
AANBOD versus VRAAG

Afnemers

	Aanbod score	Organisatie 2 Voorinvullen belastingaangifte (*jaarlijks*) Relevant	Vraag score	Verwerkbaar	Organisatie 3 Statistiek werkgelegenheid (*maandelijks*) Relevant	Vraag score	Verwerkbaar	Organisatie 4 Gratis rechtshulp? (*vraag/antwoord*) Relevant	Vraag score	Verwerkbaar
BRON: Organisatie 1		**Ja**		+	**Ja**		+	**Ja**		+
Toegankelijk (verstrekking op wettelijke basis)		Ja		+	Ja		+	Ja		+
Vertrouwenswaardig (enquête afnemers)	0,80	Ja	0,75	+	Ja	0,70	+	Ja	0,80	+
DATABASE: Inkomensgegevens		**Ja**		+	**Ja**		+	**Ja**		+
Begrijpelijk (enquête afnemers)	0,95	Ja	0,90	+	Ja	0,85	+			
Waarachtig (enquête afnemers)	0,80	Ja	0,75	+	Ja	0,70	+	Ja	0,80	+
GEGEVENSELEMENTEN:										
Betrokkene (burgerservicenummer)		**Ja**		+	Totalen		-	**Ja**		+
Variabiliteit	0	Ja	0	+				Ja	0	+
Syntactische juistheid	0,97	Ja	0,90	+				Ja	0,95	+
Werkgever/uitkeringsinstantie		**Ja**		+						
Dekkingsgraad	0,80	Ja	0,70	+						
Sectorcode (horeca, handel, industrie, etc.)					**Ja**		+			
Semantische juistheid (volgens steekproeven)	0,90				Ja	0,80	+			
Soort inkomen (loon, uitkering, etc.)					**Ja**		?			
Semantische juistheid (geen meting)	?				Ja	0,80	?			
Bedrag cumulatief inkomen		**Ja**		+				**Ja**		-
Variabiliteit	1									
Vullingsgraad	0,30	Ja	0,10	+				Ja	0,80	-
Actualiteit (duur in-/uitschrijven in dagen)	5	Ja	10	+				Ja	2	-
Tijdigheid (ouderdom, max. aantal weken)	4							Ja	0	-

Conform de NORA-principes maakt Organisatie 1 bekend over welke gegevens ze beschikt. Daartoe heeft Organisatie 1 een kwaliteitsprofiel opgesteld dat voor elk van de gegevenselementen de kwaliteitskenmerken beschrijft die Organisatie 1 daarvoor hanteert. Gegevensgerichte controles hebben geleid tot meetresultaten per kwaliteitskenmerk, deze staan in de kolom 'Aanbod score'.

In het voorbeeld is in de database de Betrokkene geregistreerd via het burgerservicenummer. Dit gegevenselement verandert niet in de tijd en daarmee is de variabiliteit 0. Een enkele keer is in de database foutief de letter 'o' in plaats van het cijfer '0' in het BSN opgenomen. Daarom blijft de syntactische juistheid van het gegevenselement Betrokkene steken op 0,97.

In de database zijn gegevens opgenomen van 80 procent van de werkgevers en uitkeringsinstanties (dekkingsgraad). Daarbij wordt tevens de Sectorcode geregistreerd. Door middel van steekproeven is bekend dat 90 procent van de geregistreerde sectorcodes semantisch juist is, dus overeenstemt met de werkelijkheid. Daarnaast bevat de database het Soort inkomen, maar hierover heeft Organisatie 1 geen meetresultaten per kwaliteitskenmerk.

Van het Bedrag cumulatief inkomen volgt de eigenaar vier kwaliteitskenmerken, namelijk de variabiliteit, de vullingsgraad, de actualiteit en de tijdigheid. De variabiliteit van het bedrag is op 1 gesteld, omdat het veelvuldig wijzigt. Het wordt maandelijks overschreven en wijzigt ook regelmatig tussendoor, bijvoorbeeld als een werknemer van werk wisselt of met pensioen gaat. In het voorbeeld is de vullingsgraad van het veld Cumulatief inkomen laag, slechts bij 30 procent van de betrokkenen is er een Bedrag cumulatief inkomen opgenomen.

Voor de tijdgerelateerde kenmerken is vastgesteld dat Organisatie 1 aan- en afmeldingen door werkgevers binnen vijf dagen verwerkt in de database. Werkgevers leveren aan het eind van elke maand de betaalde bedragen van de salarissen aan Organisatie 1 aan. Daardoor kan het bedrag voor het cumulatief inkomen vier weken oud zijn.

MAB04_binnen.indd 215 29/03/10 4:51 PM

2.3.2 Vraag

Naast dit kwaliteitsprofiel van de aanbiedende Organisatie 1, zijn ook de gevraagde profielen van Organisaties 2, 3 en 4 weergegeven. Organisatie 2 overweegt om op de jaarlijkse formulieren voor de belastingaangifte alvast het cumulatief inkomen in te vullen als service naar de burger. Daarvoor zijn de gegevenselementen Betrokkene, Werkgever en Bedrag cumulatief inkomen relevant. De door Organisatie 2 gevraagde score op de kwaliteitskenmerken ligt steeds lager dan wat Organisatie 1 aanbiedt. Daarom zijn de gegevens in de Database Inkomensgegevens van Organisatie 1 prima verwerkbaar voor Organisatie 2.
Organisatie 3 heeft als opdracht maandelijks een rapport uit te geven over de ontwikkeling van de werkgelegenheid. Daartoe heeft ze aantallen werknemers of uitkeringsgerechtigden nodig per Sectorcode en per Soort inkomen. Een aanlevering op individueel niveau zoals geregistreerd in de database Inkomensgegevens, is niet direct bruikbaar voor Organisatie 3. Hier is dus een bewerkingsslag nodig om tot totalen te komen. Met de door Organisatie 1 te leveren Sectorcode kan Organisatie 3 prima werken, de geleverde semantische juistheid is hoger dan de processen van Organisatie 3 verlangen. Echter, voor wat betreft het Soort inkomen levert Organisatie 1 geen meetresultaten betreffende de semantische juistheid. Dat terwijl Organisatie 3 wil dat deze voor 80 procent overeenstemmen met de werkelijkheid. Om de gegevens in de database toch te kunnen gebruiken, zal Organisatie 3 daarom zelf dit kenmerk moeten gaan controleren, dan wel Organisatie 1 bewegen dit voor hen te gaan doen.
In het voorbeeld is Organisatie 4 een bureau voor rechtshulp en zij wil op ad-hocbasis het cumulatief inkomen weten van klanten die aan de balie komen, om te beoordelen of zij recht hebben op gratis rechtshulp. Organisatie 4 wil bij 80 procent van deze verzoeken direct antwoord kunnen geven. Helaas kan Organisatie 1 hierbij niet helpen, immers, de vullingsgraad is slechts 30 procent. Bovendien heeft Organisatie 4 meer strikte eisen ten aanzien van actualiteit en tijdigheid. Aldus komt het aangeboden kwaliteitsprofiel van Organisatie 1 onvoldoende overeen met het door Organisatie 4 gevraagde profiel en heeft gegevensuitwisseling hier geen zin.

Op de hiervoor geschetste wijze kunnen vraag en aanbod van gegevens afgestemd worden. Het is arbeidsintensief en leidt tot uitgebreide beschrijvingen van een bericht of database. Wanneer we een auto of een computer kopen, vinden we het normaal om vele vellen met specificaties te bestuderen om te beoordelen of het product aan onze eisen voldoet. Dat ligt voor informatieproducten niet anders, ook hier is er behoefte aan uitgebreide specificaties.
De eigenaar van een gegeven is verantwoordelijk voor het kwaliteitsprofiel van de gegevens die hij aanbiedt. Wanneer helder is welke kenmerken het gewenste niveau niet halen, kan de eigenaar gericht zoeken naar de factoren die de kwaliteit bedreigen en daarbij passende maatregelen invoeren.

3 Dreigingen

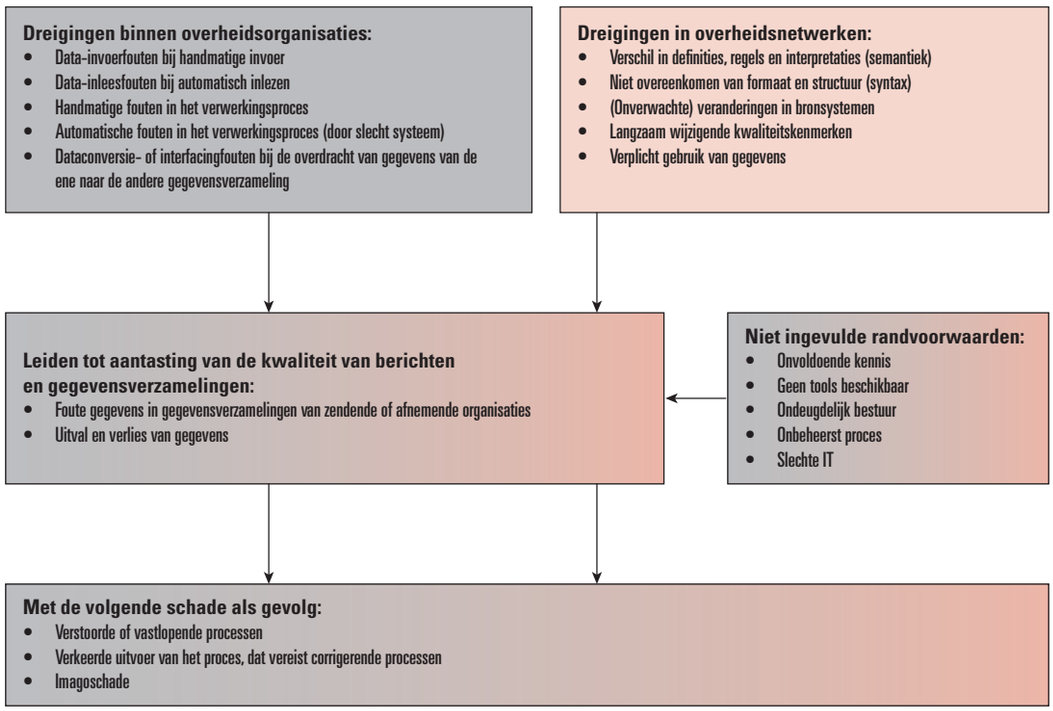
Gegevenskwaliteit is een nieuw onderzoeksgebied en er is nog geen uitgekristalliseerde systematiek om problemen aan te pakken. Om toch tot een raamwerk of normenkader te komen, is in dit onderzoek gekozen voor een benadering vanuit risicomanagement. Deze start met het onderkennen van de dreigingen. Een dreiging is de mogelijkheid dat een gebeurtenis optreedt die schade veroorzaakt. Het risico is de kans van optreden van een dreiging maal de impact. Om passende maatregelen te kunnen treffen, moet zowel de kans als de impact ingeschat worden.
Volgens Eckerson (2002) wordt de kwaliteit van gegevens het meest bedreigd door fouten bij data-invoer en fouten bij de overdracht van gegevens. Figuur 3 geeft een overzicht van dreigingen, deze zijn onderscheiden naar dreigingen binnen overheidsorganisaties en dreigingen in overheidsnetwerken. Dit overzicht is afgeleid uit de literatuur (Batini en Scannapieco, 2006; Van Besouw, 2007; Eckerson, 2002, Kroenke, 2007; McGilvray, 2008; Pipino et al., 2006; Redman, 2004, Verreck et al., 2005; Zalm en Bosch, 2007; Zwienink en Wisse, 2008) en is getoetst en aangevuld op basis van het praktijkonderzoek in de loonaangifteketen.

3.1 Dreigingen binnen overheidsorganisaties

Binnen organisaties kunnen bij handmatige data-invoer fouten gemaakt worden die leiden tot een verminderde syntactische juistheid. Voorbeelden daarvan zijn foute spellingen, cijferverwisselingen of incorrecte codes. Ook bij automatisch inlezen kunnen fouten gemaakt worden. Bijvoorbeeld als de *referential integrity checks* uit staan. Dan worden wijzigingen in de inkomende gegevens niet opgemerkt en kunnen integriteitproblemen in de database ontstaan.
In het verwerkingsproces kunnen handmatige fouten gemaakt worden. Een medewerker kan vergissingen maken of over onvoldoende kennis beschikken. Maar helaas voeren mensen soms ook doelbewust foute gegevens in of manipuleren zij gegevens in databases. Een slecht ontworpen systeem kan leiden tot automatische fouten in het verwerkingsproces van gegevens. Slecht gebouwde systemen leiden tot storingen waardoor programma's gegevens corrumperen of verkeerde uitvoer produceren. Daarnaast kan een systeem onvoldoende flexibel van opzet zijn. Bijvoorbeeld wanneer een database een vaste indeling kent die niet kan worden aangepast als nieuwe wetgeving het opslaan van een extra veld vereist.
Dataconversie- of interfacingfouten kunnen optreden bij de overdracht van gegevens van de ene naar de andere gegevensverzameling binnen een organisatie. Mogelijk

MAB04_binnen.indd 216 29/03/10 4:51 PM

Figuur 3 Dreigingen voor gegevenskwaliteit en mogelijke schade



zijn er geen datamodellen voorhanden of zijn deze niet actueel. Als ze er wel zijn, nemen programmeurs mogelijk niet de tijd om de datamodellen van het bestaande systeem en het doelsysteem te begrijpen, waardoor ze software ontwikkelen die fouten introduceert. Eén fout in een datamigratieprogramma of in een interface kan leiden tot fouten in duizenden regels.

3.2 Dreigingen in overheidsnetwerken

Verschillende organisaties werken vaak met licht afwijkende definities voor begrippen, interpreteren begrippen anders en berekenen waarden volgens andere regels. Naast deze semantische verschillen kan ook de syntax afwijken. Het formaat of de structuur kan bijvoorbeeld anders zijn in naam (eerst voornaam of eerst achternaam), in het gegevensformaat (zes byte datavelden versus vier byte datavelden) of in codering (man-vrouw versus m-f versus 1-2). Veranderingen in bronsystemen, vooral onverwachte, kunnen leiden tot problemen bij afnemers. Dat kan bijvoorbeeld gebeuren als de kwaliteitscontrole bij de bronhouder tijdelijk niet operationeel is. Onderzoek laat zien dat maandelijks 2 procent van een klantenbestand wijzigt doordat mensen verhuizen, trouwen, scheiden en overlijden (Eckerson, 2002). Wanneer de kenmerken van gegevens van de bron geleidelijk wijzigen, kan er een kwaliteitsissue ontstaan bij de afnemer van de gegevens. Wanneer de brongegevens fouten bevatten en er een wettelijke gebruiksplicht geldt, raken mogelijk ook de gegevensverzamelingen van de afnemers vervuild.

3.3 Niet-ingevulde randvoorwaarden

Naast de directe dreigingen kunnen ook niet-ingevulde randvoorwaarden een negatieve invloed hebben op de kwaliteit van informatieproducten. Gegevenskwaliteit is ingewikkelde materie, zowel kennis als tools zijn nodig. Tools zijn geautomatiseerde procedures, bijvoorbeeld voor objectidentificatie, data-integratie of het lokaliseren van fouten. Tools zijn nodig om efficiënt te kunnen werken, maar vereisen een grondig begrip van het werkveld. Naast kennis en tools zijn procesbeheersing, een adequate ondersteuning van het proces door IT en goed bestuur randvoorwaarden.

3.4 Gevolgen

De genoemde bedreigingen en niet-ingevulde randvoorwaarden kunnen leiden tot foute gegevens in berichten en databases en tot uitval en verlies van gegevens. Een vervuilde gegevensverzameling bevat mogelijk inconsistente waarden, foutief gespelde namen of verouderde of onmogelijke waarden. Deze aantasting van de gegevenskwaliteit leidt tot schade. Processen raken verstoord of lopen vast, dat geeft wrijving met de afnemers van de producten uit dat proces. Wanneer de fouten in gegevens niet ontdekt worden, kan het leiden tot verkeerde uitvoer van het proces. Daarvoor moeten dan weer kost-

MAB04_binnen.indd 217 29/03/10 4:51 PM

bare processen ingericht worden die deze fouten corrigeren. Redman (2004) ontwikkelde door de jaren heen de *Rule of Ten: If it costs 1.00 dollar to complete a simple operation when all the data is perfect, then it costs 10.00 dollar when it is not (i.e., late, hard to interpret, incorrect, etc.).*
In de massale processen van bijvoorbeeld de Belastingdienst leiden ontbrekende gegevens tot uitval die handmatig behandeld moet worden. Dat is heel arbeidsintensief. In het ergste geval genereren foute gegevens foute beschikkingen aan burgers en bedrijven. Dat leidt tot bezwaarprocedures en imagoschade.

4 Verbetermogelijkheden

Voor het beheersen van de kwaliteit bij het uitwisselen van gegevens in overheidsnetwerken staat een scala aan maatregelen ter beschikking. Tabel 2 geeft een overzicht dat afgeleid is uit diverse literatuur (Batini en Scannapieco, 2006; Van Besouw, 2007; De Bruijn et al., 2006a en 2006b; Eckerson, 2002; English, 2002; Kroenke, 2007; McGilvray, 2008; Pipino et al., 2006; Verreck et al., 2005; Zalm en Bosch, 2007) en getoetst en aangevuld is op basis van het praktijkonderzoek in de loonaangifteketen. Daarbij worden de verbetermogelijkheden binnen organisaties onderscheiden van de mogelijkheden bij gegevensuitwisseling in netwerken.

Tabel 2 Mogelijkheden om de kwaliteit van gegevens te bevorderen

Verbetermogelijkheden binnen overheidsorganisaties
Gegevensgericht
Procesgericht
Managementcyclus
Verbetermogelijkheden in overheidsnetwerken
Wetgeving
Afspraken
Kwaliteitslabels
Ondersteunen toeleveranciers van gegevens

4.1 Verbetermogelijkheden binnen organisaties

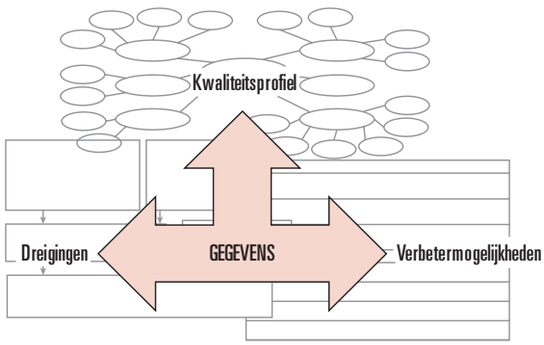
Overheidsorganisaties kunnen intern de kwaliteit van gegevens verbeteren door een gegevensgerichte aanpak, door het verbeteren van de verwerkingsprocessen of door het uitbouwen van beide tot een managementcyclus.
Een gegevensgerichte aanpak bevat bijvoorbeeld controles op het voldoen aan bedrijfsregels en op de aanwezigheid van gegevens van de benodigde populatie. De mate van overeenstemming met de werkelijkheid is te bevorderen door objectidentificatie, data-analyse en door correctie op basis van waarneming.
Bij fundamentele problemen is herontwerp van de processen nodig, daarbij kunnen gegevensgerichte controles ingevoerd worden. Een goed ontworpen change management-proces bijvoorbeeld voorkomt het introduceren van fouten wanneer processen of systemen wijzigen.
Het inrichten van een managementcyclus voor gegevenskwaliteit brengt beide benaderingen bij elkaar. De start van een managementcyclus is een risicoanalyse. Organisatorische maatregelen als het helder beleggen van het eigenaarschap en het motiveren van medewerkers voor een goede kwaliteit van gegevens dragen eveneens bij. Audits op zowel de kwaliteit van gegevens als op het proces kunnen richting geven aan verbeteringen.

4.2 Verbeteren in overheidsnetwerken

In overheidsnetwerken heeft wetgeving, het maken van afspraken, het werken met kwaliteitslabels en het ondersteunen van toeleveranciers een positieve invloed op de kwaliteit van uitgewisselde berichten.
Wetgeving is het zwaarste middel om de kwaliteit van gegevens te beheersen. De wet kan authentieke gegevens definiëren en het slechts eenmaal uitvragen en meervoudig gebruik ervan afdwingen. Het verplicht terugmelden van fouten in de gegevens door afnemers leidt tot een betere kwaliteit bij de bronhouder. Tevens kan de wet burgers en bedrijven verplichten om gegevens aan te leveren.
Een andere verbetermogelijkheid ligt in het maken van afspraken tussen partners, al dan niet in een sector. Dat kan gaan over definities, over gebruik of over het doorgeven van wijzigingen door de zender en het terugmelden van fouten door afnemers. Het vastleggen van de kwaliteitsprofielen van zender en afnemer maakt gerichte monitoring en verbetering mogelijk.
In netwerken helpen de beschikbaarheid van meetresultaten per kwaliteitskenmerk en van metagegevens over de afkomst van de gegevens. Deze kunnen vastgelegd worden met kwaliteitslabels. Ze worden bijvoorkeur in het bericht opgenomen, zodat iedere afnemer zelf kan beoordelen of het bericht voor zijn toepassing bruikbaar is. Vergelijk het met een etiket in een kledingstuk. Dat beschrijft het materiaal en hoe met het kledingstuk omgegaan moet worden. De verantwoordelijkheid voor het correct gebruik van de ontvangen gegevens ligt bij de afnemer. Voor situaties waarin de afnemer de kwaliteit van gegevens niet zelf kan vaststellen, behoort een mededeling van de afzender of een certificaat door een derde partij tot de mogelijkheden (De Bruijn et al., 2006a en 2006b). Een technologische oplossing is een programma dat in een netwerk de kwaliteit van circulerende berichten meet.
Burgers en bedrijven die gegevens toeleveren, kunnen ondersteund worden door het vereiste kwaliteitsprofiel helder te communiceren en door het bieden van een helpdesk en testfaciliteiten. Daarmee kan vervuiling van databases voor een deel voorkomen worden.

MAB04_binnen.indd 218 29/03/10 4:51 PM

Figuur 4 Datadriehoek, verbeteren van de kwaliteit van gegevens met drie invalshoeken



5 Datadriehoek

5.1 Model en gebruik

De voorgaande paragrafen hebben de problematiek van gegevenskwaliteit benaderd vanuit de invalshoeken kwaliteitsprofiel, dreigingen en verbetermogelijkheden. Gecombineerd vormen deze drie invalshoeken het managementmodel voor gegevenskwaliteit de 'Datadriehoek', dat afgebeeld is in figuur 4.

De Datadriehoek is primair gericht op het management dat invulling wil geven aan de verantwoordelijkheid voor de kwaliteit van zijn gegevens. Het model helpt adviseurs bij het analyseren van problemen met de kwaliteit van gegevens en bij het definiëren van de huidige en de gewenste situatie. Voor auditors kan het een aanvulling vormen op de bestaande normenkaders.
Het management begint met het definiëren van het vereiste kwaliteitsprofiel. Op basis van de eisen van afnemende processen stelt de eigenaar van de gegevens vast welke gegevenselementen hij aanbiedt, welke kwaliteitskenmerken daarvoor gelden, wat de streefwaarde per kenmerk is en binnen welke toleranties deze waarde moet liggen. Dat is beschreven in paragraaf 2.
Daarna volgt het identificeren van de dreigingen. Als helder is waar het mis gaat met de gegevens, kiest het management op basis van deze dreigingen en het gedefinieerde kwaliteitsprofiel een passende mix van maatregelen uit de lijst van verbetermogelijkheden. Zie hiervoor paragraaf 3 en paragraaf 4. Na het invoeren van de maatregelen kan het management bijsturen op basis van de gemeten waarden van de kwaliteitskenmerken.

5.2 Toepasbaarheid

De Datadriehoek helpt adviseurs bij het analyseren van problemen en bij het definiëren van de IST- en de SOLL-positie. De huidige situatie met betrekking tot gegevenskwaliteit in een proces of een organisatie kan eenvoudig op de raamwerken per invalshoek geplot worden. Dat maakt duidelijk wat er op dit moment geregeld is. De niet-ingevulde delen kunnen de adviseurs op ideeën brengen bij het definiëren van de gewenste situatie. In de praktijkcasus van de loonaangifteketen bleek het model een handige werkwijze te zijn, die snel zicht geeft op problemen en oplossingen aanreikt.

Een audit naar de beheersing van de kwaliteit van gegevens kan meerdere vormen hebben. Het object van de audit kan een product of een proces zijn. Een productgerichte audit onderzoekt bijvoorbeeld of een bericht aan de streefwaarden per kwaliteitskenmerk voldoet. Een procesgerichte audit onderzoekt bijvoorbeeld het synchronisatieproces tussen twee databases of het change management-proces. Een combinatie van een product- en een procesaudit kan ook. Bij een managementinformatiesysteem bijvoorbeeld kan zowel de inhoud van de gegevens, de applicatie zelf als het beveiligingsproces om het systeem heen onderzocht worden.
Daarnaast wordt bij audits onderscheid gemaakt naar onderzoeken die zich richten op een ontwerp, op een effectieve werking, op een diagnose of op het signaleren van een probleem. Een ontwerpgerichte audit onderzoekt bijvoorbeeld of de kwaliteitsprofielen eenduidig beschreven zijn. Ook kan het ontwerp van het proces onderzocht worden. Zijn de stappen helder omschreven en zitten er voldoende controlemomenten in? Is de mix van getroffen maatregelen evenwichtig en gericht op de onderkende dreigingen en kritieke kwaliteitskenmerken?
Een audit gericht op bestaan en werking onderzoekt of de ontworpen maatregelen ook daadwerkelijk geïmplementeerd zijn en effectief werken. Worden berichten inderdaad eenduidig geïnterpreteerd en zijn de meetwaarden van de belangrijkste kwaliteitskenmerken bekend? Ook kunnen er totaalcontroles uitgevoerd worden. Bijvoorbeeld of de totalen van de inkomensgegevens van werkgevers overeenkomen met de totalen van de opgaven die de werknemers individueel gedaan hebben. Andere audits gericht op het geven van zekerheid onderzoeken bijvoorbeeld de werking van een managementcyclus voor gegevenskwaliteit binnen een organisatie of het functioneren van de gegevensuitwisseling in een netwerk.
Ook kan een audit diagnostisch van aard zijn, bijvoorbeeld bij grote uitval van gegevens in een massaal proces. Dan zal vooral het raamwerk met dreigingen van de Datadriehoek helpen. Ten slotte zijn er de probleemsignalerende audits. Bijvoorbeeld als de doelmatigheid van arbeidsintensieve controles ter discussie staat.

6 Conclusies

De elektronische overheid ontwikkelt zich tot een netwerk van organisaties die continu gegevens uitwisselen. De kwaliteit van deze berichten bepaalt de mogelijkheid om

MAB04_binnen.indd 219 29/03/10 4:51 PM

ze te verwerken bij afnemers. Daarom is het nodig om met de kwaliteit van gegevens aan de slag te gaan. Het managementmodel de Datadriehoek helpt daarbij door het aanbieden van raamwerken voor kwaliteitskenmerken, dreigingen en verbetermogelijkheden.

In netwerken worden elektronische berichten zelfstandige producten met eigen specificaties. Berichten en databases zijn te beschrijven door de gegevenselementen die ze bevatten met hun kwaliteitskenmerken. De bekende kwaliteitscriteria juistheid, volledigheid en tijdigheid worden voor gegevens verbijzonderd. De juistheid van gegevens wordt bepaald door de syntactische en semantische nauwkeurigheid en door het voldoen aan bedrijfs- en integriteitregels. Volledigheid is de resultante van de dekkingsgraad en de mate waarin een gegevensverzameling compleet is. Voor afnemers is de tijdige beschikbaarheid van actuele gegevens belangrijk.

Bij uitwisseling van gegevens spelen additionele netwerkgerelateerde kwaliteitskenmerken. Daaronder valt het toegankelijk zijn van een bron en het kunnen vertrouwen op de gegevensleveringen van de desbetreffende organisatie. Het schema van een bericht of de gegevensverzameling waaruit het afkomstig is, moet goed in elkaar zitten. Een afnemer hecht waarde aan een begrijpelijk, waarachtig en relevant bericht dat te verwerken is door zijn processen. De mate waarin gegevens variabel zijn, speelt vooral als meerdere gegevensverzamelingen synchroon gehouden moeten worden.

Het kwaliteitsprofiel van een bericht of database specificeert de kwaliteitskenmerken per gegevenselement, de streefwaarde per kenmerk en de tolerantie die daarbij wordt toegestaan. Een afnemer beoordeelt of een beschikbare gegevensset voor hem bruikbaar is door zijn kwaliteitsprofiel naast het aangeboden profiel te leggen.

Binnen overheidsorganisaties bedreigen fouten bij de invoer, de verwerking en de conversie van gegevens de kwaliteit. In overheidsnetwerken vormen afwijkingen in semantiek en syntax, veranderende bronsystemen en verplicht gebruik de dreigingen. Daarnaast hebben nietingevulde randvoorwaarden zoals een onbeheerst proces of het ontbreken van tools een negatieve invloed op de kwaliteit. Dat leidt tot foute gegevens, uitval en verlies van gegevens, tot verstoorde processen met verkeerde uitvoer en tot imagoschade.

De kwaliteit van gegevens kan binnen overheidsorganisaties verbeterd worden door een gegevensgerichte aanpak, door het verbeteren van het proces of door het uitbouwen van beide tot een managementcyclus. In overheidsnetwerken bevordert wetgeving, het maken van afspraken, het werken met kwaliteitslabels en het ondersteunen van toeleveranciers de kwaliteit van gegevens. De kwaliteit van gegevens kan door het herontwerpen van processen verbeteren, maar dat is niet vanzelfsprekend het gevolg. Daarvoor zijn aanvullende verbeteringen nodig. Zowel productgerichte als procesgerichte audits kunnen zekerheid geven over de mate van beheersing van de gegevenskwaliteit binnen organisaties of in netwerken.

Het delen en uitwisselen van gegevens is nodig om de dienstverlening van de overheid te kunnen verbeteren. Wat we samen kunnen bereiken, is een verlaging van de kosten die gemoeid zijn met het oplossen van problemen door onjuiste gegevens en een toename van de inkomsten. Door meer integraal te werken en gegevens te combineren, kan de overheid fraude effectiever bestrijden. Maar bovenal gaat het om een verbeterde dienstverlening en het vertrouwen van burgers en bedrijven dat hun gegevens bij de overheid in goede handen zijn. ■

Mw. ir. C.W.M. (Carolien) Besselink EMITA geeft binnen het managementteam Belastingdienst/Centrale Administratie leiding aan de units Gegevensdiensten en Informatiemanagement. Zij is daarnaast verbonden aan de opleiding Executive Master of IT Auditing van de Universiteit van Amsterdam. Daarvoor bekleedde zij verschillende functies op het raakvlak van business en IT in de publieke en private sector.

MAB04_binnen.indd 220 29/03/10 4:51 PM

Literatuur

- Batini, C. en M. Scannapieco (2006), *Data quality, concepts, methodologies and techniques*, Springer.
- Besouw, van, F. (2007), De samenhang tussen bedrijfsprocessen, bedrijfsregels en gegevenskwaliteit, *Informatie*, juni, pp. 12-17; http://www.informatie.nl/Artikelen/2007/juni/Default.aspx.
- Besselink, C.W.M. (2009), *Gegevenskwaliteit in overheidsnetwerken*; http://im.fee.uva.nl/riga/.
- Bruijn, de, A.J.M., A.J.M. van der Meer, M.C.M. Slot, P.C.J. Nieuwenhuizen en B.J. van Staveren (2006a), Ketengovernance: startpunt voor keteninrichting en ketenauditing, *EDP-auditor*, no. 1, pp. 28-37.
- Bruijn, de A.J.M., A.J.M. van der Meer, M.C.M. Slot, P.C.J. Nieuwenhuizen en B.J. van Staveren (2006b), Ketengovernance, ketensamenwerking binnen het publieke domein, *EDP-auditor*, no. 2, pp. 41-46.
- *Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the Re-use of Public Sector Information* (2003), European Parliament; http://ec.europa.eu/information_society/policy/psi/docs/pdfs/directive/psi_directive_en.pdf.
- Donner, J.P.H. en J.C. de Jager (29 april 2008), *Normenkader voor de werkende loonaangifteketen*, brief aan de voorzitter van de Tweede Kamer der Staten-Generaal; http://docs.minszw.nl/pdf/129/2008/129_2008_3_11762.pdf.
- Eckerson, W.W. (2002), *Data warehousing special report: Data quality and the bottom line*; http://adtmag.com/articles/2002/05/01/data-warehousing-special-report-data-quality-and-the-bottom-line.aspx.
- English, L. (2002), The essentials of information quality management, *Information Management Magazine*; http://www.information-management.com/issues/20020901/5690-1.html.
- *European Interoperability Framework* (2004), European Communities; http://ec.europa.eu/idabc/en/document/3473.
- *Interoperabiliteitsagenda, Ruimte voor richting* (2008), Forum Standaardisatie; http://www.forumstandaardisatie.nl/fileadmin/OVOS/bijlage_bij_CS04-11-06A_Interoperabiliteitsagenda.pdf.
- *IPA, Integrale probleemanalyse loonaangifteketen* (12 september 2007), UWV en Belastingdienst; http://docs.minszw.nl/pdf/35/2007/35_2007_3_10928.pdf.
- ISO/DIS 8000 (2008) Data quality - Part 102: *Master data: Exchange of characteristic data: Vocabulary* en Part 110: *Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance tot data specification*, International Organization for Standardization. Deze twee ontwerpnormen vervangen ISO/TS 8000-110:2008.
- McGilvray, D. (2008), *Executing data quality projects, Ten steps to quality data and trusted information*, Morgan Kaufmann Publishers.
- Kroenke, D.M. (2007), *Databases*, 10e druk, Pearson Education Benelux.
- *NORA, Nederlandse Overheid Referentie Architectuur 2.0* (2007), ICTU; https://www.surfgroepen.nl/sites/NORA-architecten/Webpaginas/Over%20NORA.aspx.
- NOREA Geschrift No. 1 (1998), *IT-auditing aangeduid.*
- *Normenkader werkende loonaangifteketen* (25 april 2008), UWV en Belastingdienst; http://docs.minszw.nl/pdf/129/2008/129_2008_3_11762.pdf.
- Pipino, L.L., J.D. Funk en R.Y. Wang (2006), *Journey to data quality*, The MIT Press.
- Redman, T.C. (2004), Data: An unfolding quality disaster, *Information Management Magazine*; http://www.information-management.com/issues/20040801/1007211-1.html.
- Smits, L.J.E. (12 november 2008), *Oordeel negende halfjaarlijkse rapportage Samenwerking UWV Belastingdienst (SUB)*, brief aan minister van SZW en de staatssecretaris van Financiën, HEC; http://static.ikregeer.nl/pdf/BLG20256.pdf.
- Starreveld, R.W. (2006), *Bestuurlijke informatieverzorging, deel 2B Toepassingen*, 6e druk, Wolters-Noordhoff.
- *Sturing op de kwaliteit van gegevens* (27 april 2008), brief aan minister van SZW en de staatssecretaris van Financiën, Ketenmanager; http://docs.minszw.nl/pdf/35/2008/35_2008_3_11765.pdf.
- Verreck, O., A. de Graaf en W. van der Sanden (2005), Meten en verbeteren van gegevenskwaliteit, *.ego*, vol. 5, no. 1, pp. 22-25; http://www.sbit.nl/ego/bestanden/8Meten%20en%20verbeteren%20van%20gegevenskwaliteit.pdf.
- Zalm, van der, M. en J. van den Bosch (2007), Van accountantsproof naar bedrijfsvoeringsproof, *Tijdschrift Controlling*, augustus, pp. 30-34; http://www.vka.nl/sites/default/files/downloads/30-34_ZalmBosch_7.8_2007.pdf.
- Zwienink, S. en P. Wisse (2008), *Eerlijk zullen we alles delen, verkenningen naar interoperabiliteit*, Forum Standaardisatie; http://gbo.overheid.nl/nieuws/artikel/197.
- *Zorgen over kwaliteit digitale data overheid* (24 april 2009), Automatiseringsgids.

MAB04_binnen.indd 221 29/03/10 4:51 PM