Werken met centraal configuratiesheet bespaart ontwikkeltijd

Het ontsluiten van (externe) niet-gestructureerde gegevens

Sjoerd Hobo en Rob Peters

Data-integratie is veelal gericht op gestructureerde gegevens. Dat komt omdat die gegevens uit gestructureerde systemen, zoals databases, van binnen de organisatie komen. Indien de gegevens uit een niet-gestructureerde interne bron komen, dan kan men een structuur afdwingen die aansluit bij de overige gegevens. Bijvoorbeeld door het gebruik van reeds eerder gebruikte datatypes en codes.

Organisaties krijgen meer en meer de mogelijkheid om uit externe bronnen gegevens toe te voegen aan die interne gestructureerde gegevens. Dit is een verrijking, omdat zo bijvoorbeeld een productiebedrijf productiegegevens kan koppelen aan detail-handel-verkoopgegevens, verkregen van de grossier. Deze externe gegevens zijn regelmatig niet gestructureerd. In dit artikel staan we stil bij de uitdagingen die het ontsluiten van externe en niet gestructureerde gegevens met zich meebrengen.

Schijnnauwkeurigheid

Externe gegevens komen van een derde partij of leverancier (dit kan overigens wel een organisatie zijn die onderdeel uitmaakt van de eigen organisatiestructuur). Daarom is het moeilijk tot onmogelijk invloed uit te oefenen op de aangeleverde externe gegevens. Het komt zoals het komt. Hierdoor wordt het ook een stuk moeilijker de gegevens te begrijpen. Zo kan een extern bestand een kolom winkelvoorraad bevatten, maar is dit nu inclusief of exclusief de display-voorraad? Soms worden separaat de definities (metadata) van de elementen in het externe gegevensbestand geleverd. Echter deze definities zijn moeilijk te controleren. Het komt voor dat regelmatig met de aanleverende partij moet worden overlegd voordat alle definities en uitzonderingen duidelijk zijn. Naast deze definitieproblematiek, kan ook de kwaliteit en de consistentie van de gegevens te wensen overlaten. Natuurlijk kan de aanleverende partij erop gewezen worden dat er fouten in het externe bestand zitten, maar vaak komt het er toch op neer dat zelf (intern) maatregelen getroffen moeten worden om deze kwaliteit-issues het hoofd te bieden. Zowel metadata als gegevenskwaliteit-issues maken dat er goed moet worden nagegaan wat de betrouwbaarheid is van het externe bestand.

Te vaak constateren ondergetekenden in de praktijk dat er blindelings op externe gegevens wordt gestuurd, terwijl het schijnnauwkeurigheid betreft. Trend-analyse zou meer op zijn plaats zijn.

Ook de structuur van de externe gegevens speelt een belangrijke rol. Indien de externe gegevens gestructureerd worden aangeleverd dan laten die zich makkelijk extraheren, transformeren en laden in bestaande gegevensomgevingen (zoals bijvoorbeeld een datawarehouse). Dat wordt moeilijk wanneer externe gegevens niet gestructureerd worden aangeleverd. Dit kan het geval zijn omdat de externe gegevens in PDF-formaat worden aangeleverd en/of in een formaat dat sterk afwijkt van de gewenste tabellen met zijn kolommen of XML/XSD layout.

Een gestructureerd gegevensbestand geniet de voorkeur, omdat

Microsoft Excel - Cust_CFG_DDI.xls

File Edit View Insert Format Tools Data Window Help

	A	B	C	D
1	CUST_ID	TYPE_ELEMENT	ELEMENT	VALUE
2	CC.DE.SAP.0002023154	SRC_CHECK	D1	LAGER
3	CC.DE.SAP.0002023154	SRC_EXT	SRC_EXT	XLS
4	CC.DE.SAP.0002023154	HDR_SIZE	HDR_SIZE	2
5	CC.DE.SAP.0002023154	DATE_MASK	DATE_MASK	YYYY-MM
6	CC.DE.SAP.0002023154	SHT_ACT	SHT_ACT	1
7	CC.DE.SAP.0002023154	FIX_VALUE	SRC_TYPE	SOD0
8	CC.DE.SAP.0002023154	SRC_SEP	SRC_SEP	
9	CC.DE.SAP.0002023154	SRC_DLM	SRC_DLM	
10	CC.DE.SAP.0002023154	FIX_VALUE	TTYPE	SOD0000001
11	CC.DE.SAP.0002023154	FIX_VALUE	VERSION	1.0
12	CC.DE.SAP.0002023154	FIX_VALUE	NSO	DE
13	CC.DE.SAP.0002023154	FIX_VALUE	DATA_PROV_TYPE	SAP
14	CC.DE.SAP.0002023154	FIX_VALUE	DATA_PROV_ID	0002023154
15	CC.DE.SAP.0002023154	FIX_VALUE	CREATOR_NAME	N.A.
16	CC.DE.SAP.0002023154	FIX_VALUE	SAP_CUST_NO	0002023154
17	CC.DE.SAP.0002023154	FIX_VALUE	CC	DE
18	CC.DE.SAP.0002023154	FIX_VALUE	COUNT_SHOPS	6
19	CC.DE.SAP.0002023154	SRC_CELL	START_DATE	E1
20	CC.DE.SAP.0002023154	SRC_COL	CTN	1
21	CC.DE.SAP.0002023154	SRC_COL	ART_DESC	2
22	CC.DE.SAP.0002023154	SRC_COL	SLS_OUT	3
23	CC.DE.SAP.0002023154	SRC_COL	TOT_STK	4
24	CC.DE.OPCO.N96	SRC_CHECK	L8	ONAT/JAHR
25	CC.DE.OPCO.N96	SRC_EXT	SRC_EXT	XLS
26	CC.DE.OPCO.N96	HDR_SIZE	HDR_SIZE	9
27	CC.DE.OPCO.N96	DATE_MASK	DATE_MASK	YYYY-MM

Cust_CFG_DDI | Xtra Dynamic Lookup

Afbeelding 1: Een configuratiebestand wordt gebruikt bij het structuren (in kaart brengen) van externe ongestructureerde gegevens.

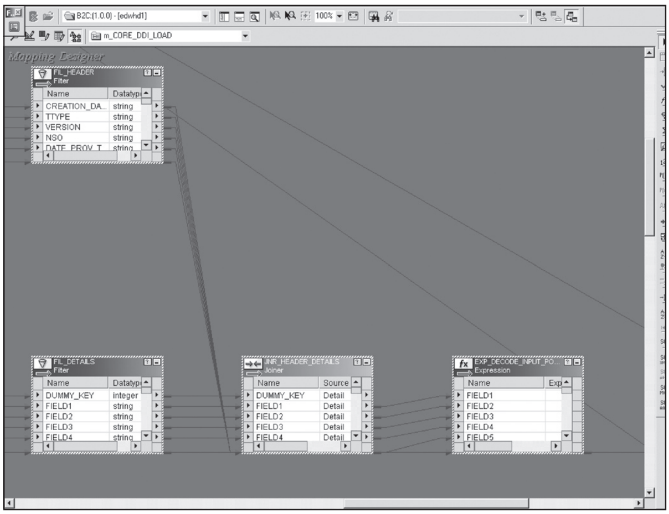
de hulpmiddelen die voorhanden zijn daar nu eenmaal goed mee uit de voeten kunnen (zoals ETL-tools, SQL). Echter, ook niet-gestructureerde gegevens worden met de hedendaagse technologie ontsloten (bijvoorbeeld met XSLT of .NET). Maar ook in de ontsluiting van deze niet-gestructureerde gegevens is flexibiliteit, snel kunnen inspelen op veranderingen en uitbreidingen, van belang. Daarom verdient ook hier een ETL-tool de voorkeur boven het programmeren van een oplossing.

Mogelijke inrichtingen

Afhankelijk van de situatie kun je als ontvangende partij invloed uitoefenen op het proces en de vorm van aanlevering van gegevens door de externe partij.

Laat de externe partijen de interface zelf ontwikkelen.
Niets is efficiënter dan de externe gegevens kant en klaar te ontvangen, dat wil zeggen dat de gegevens eenvoudig kunnen worden geïntegreerd met de interne gegevens binnen een bestaande omgeving. Om dit doel te bereiken, zal de ontvangende partij minimaal een generiek interface moeten beschrijven waarin uitspraken worden gedaan over onder andere de gegevenselementen, de structuur, de vorm (TXT, XML, XLS) en de regelmaat. Vervolgens kan de externe partij/leverancier aan de slag met deze beschrijving om de interface technisch te realiseren. Nadeel van dit scenario is dat het proces van ontwikkeling van het interface buiten de controle valt van de organisatie die de gegevens graag wenst te ontvangen. De projectkalender van de externe partij, in dit geval de key account, kan prioriteiten bevatten die haaks staan op datgene wat de ontvangende organisatie graag zou willen zien, namelijk het bouwen en dus ontvangen van de interface met externe gegevens. Invloed uitoefenen op een externe partij is een moeilijke zaak, al helemaal als er geen win-win situatie is. Als de aanleverende partij geen gegevens of andere diensten terug ontvangt dan zal deze niet enthousiast ingaan op het verzoek tot leveren van zijn gegevens. Tevens kan de aanleverende partij bij meer producenten te boek staan als key account en dus met meer verzoeken tot het leveren van gegevens te maken hebben.

Accepteer de externe gegevens zoals ze zijn.
Om een aantal redenen kan ervoor gekozen worden externe partijen de gewenste gegevens te laten aanleveren in het beschikbare of door hun gewenste formaat. Indien de externe partij geen baat heeft bij het sturen van zijn gegevens, dan mag je als vragende organisatie al blij zijn überhaupt iets te mogen ontvangen. Zoals gesteld, kunnen distributeurs van producten door meerdere producenten worden gevraagd gegevens aan te leveren. Ook al zouden de distributeurs de kosten van de ontwikkeling van het interface terug ontvangen, dan nog is het goed denkbaar dat ze weigeren de interface te bouwen en besluiten de gegevens aan te leveren in het formaat zoals ze dat zelf kiezen. Geen enkele producent zit immers te wachten op het onderhouden en beheren van meerdere interfaces naar diverse partijen. Kortom, vaak is men gedwongen externe gegevens te ontvangen zoals aangeboden.



Afbeelding 2: Informatica PowerCenter-mapping, waarin het configuratiebestand als lookup is opgenomen (niet zichtbaar) en veelvuldig wordt aangeroepen in de expressie DECODE_INPUT_PORTS.

De uitdagingen

Het integreren van externe gegevens vergroot de bekende uitdagingen. Er wordt stilgestaan bij een drietal aandachtsgebieden: de kwaliteit van gegevens, de betekenis van de gegevens en de uitdaging van het integreren van externe gegevens aan de interne gegevens.

De kwaliteit van de gegevens.
De kwaliteit van gegevens bewaken is altijd een uitdaging, voor wat betreft zowel interne als externe gegevens. Het definiëren van business-regels en validatieregels kan voorkomen dat een datawarehouse-omgeving vervuild raakt. Mochten desondanks zich niet juiste gegevens in de datawarehouse-omgeving bevinden, kunnen deze nog altijd met behulp van een Data Quality tool boven water worden gehaald en, indien mogelijk en gewenst, gecorrigeerd worden. Interne gegevens hebben het voordeel dat ze gecorrigeerd zouden kunnen worden aan de gegevensbron kant. Een dergelijke correctie geniet de voorkeur boven correctie achteraf.
De bron van externe gegevens bevindt zich buiten het bereik van de ontvangende partij. Een kwaliteitsslag door correctie aan de bronkant is niet mogelijk. Nog belangrijker wordt het dus strikte regels te definiëren teneinde de externe gegevens te kunnen valideren. Vreemde records moeten worden afgekeurd en geparkeerd in een quarantaine-omgeving. Vervolgens moeten deze gegevens separaat worden beoordeeld en gecorrigeerd indien mogelijk, zodat de gegevens alsnog kunnen worden opgenomen in het datawarehouse of definitief worden afgekeurd. Uiteraard dient de eindgebruiker op de hoogte gesteld worden van het feit dat er externe gegevens zijn aangepast of verwijderd.

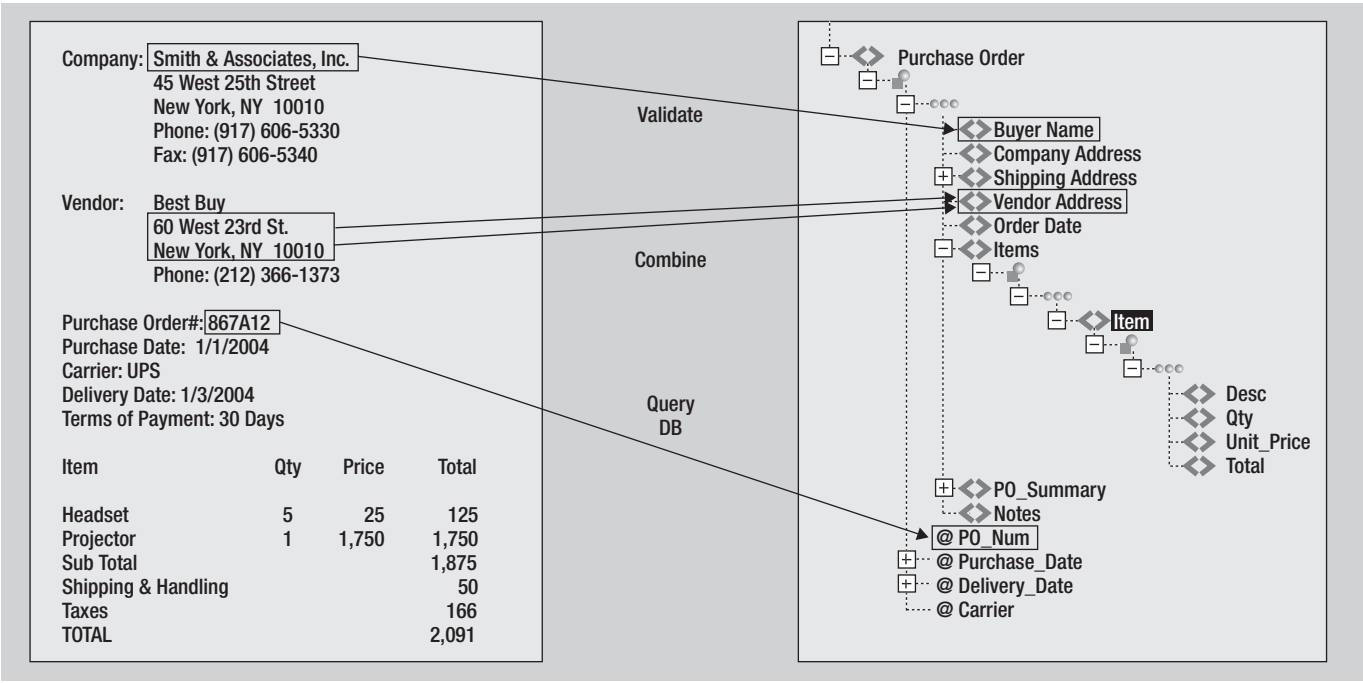
De betekenis van de gegevens.
Gegevens in het algemeen, maar externe gegevens in het bijzon-

der, kunnen er op het eerste gezicht goed uitzien maar dat zegt niets over de betekenis ervan. Het gegeven sales value ex. VAT verklapt niets over het feit of deze waarde nu inclusief of exclusief eventuele retouren is. Onbewerkte externe gegevens zijn waardevoller dan berekende externe gegevens. (Immers hoe zijn de berekende gegevens precies tot stand gekomen?) Maar ook de betekenis van onbewerkte gegevens is niet altijd even voor de hand liggend. Blindelings aannemen dat als een gegeven binnen een range van het aanvaardbare ligt, het een juist gegeven betreft is gevaarlijk. Mede daarom moet een eindgebruiker ten alle tijden worden geïnformeerd over het feit dat er naar externe gegevens wordt gekeken. Externe gegevens kunnen in sommige gevallen beter als indicatie worden beschouwd dan als de absolute waarheid.

De relatie leggen tussen externe en interne gegevens.
De toegevoegde waarde van de externe gegevens wordt verkregen door koppeling aan de interne gegevens. Bijvoorbeeld grossiers, de business-partners van een productiebedrijf, leveren gegevens met betrekking tot productverkopen aan de detailhandel. Deze gegevens krijgen een meerwaarde indien gekoppeld aan eigen bedrijfsgegevens, zoals de productievolumes per product. Door gegevens van aanleverende en afnemende partijen te integreren met de eigen bedrijfsgegevens, kan de supply chain in kaart worden gebracht. De koppeling van de gegevens zal zich richten op entiteiten zoals producten, klanten en leveranciers. De koppeling is eenvoudig wanneer de gegevensaanleverende partij en de ontvangende partij dezelfde algemene codes gebruiken, zoals de EAN-code voor producten of een BTW-nummer voor klanten. Echter, sommige bedrijven omzeilen een administratieve rompslomp door deze EAN-code te hergebruiken nadat een product uit productie en uit voorraad is gegaan. Lastiger wordt het wanneer beide partijen eigen, onderling niet matchende, codes gebruiken. Dan zit er niets anders op dan de producten of klanten te koppelen door vergelijking van omschrijving of NAW-gegevens. De betere datakwaliteit-tools zijn steeds beter in staat dit proces van koppeling te automatiseren. Soms zijn dergelijke tools geïntegreerd in de ETL-tool (zoals Trillium in Informatica PowerCenter). Dat is belangrijk omdat het een activiteit is die herhaald moet worden. Indien men niet de beschikking heeft over een dergelijke datakwaliteit-tool dan is herhaald een handmatige koppeling nodig. Deze activiteit moet goed belegd worden in de organisatie, omdat anders de betrouwbaarheid en daarmee de bruikbaarheid van de gegevens snel afneemt.

Een oplossing

Nu wordt een oplossing beschreven voor het ontsluiten en integreren van externe niet gestructureerde gegevens met behulp van een ETL-tool (in dit geval Informatica PowerCenter) in het slechtst denkbare scenario: iedere aanleverende partij stuurt zijn verkoop- en voorraadgegevens ieder op zijn eigen wijze. De ene partij meer gestructureerd dan de andere partij. Tot op heden zijn de meeste ETL-tools krachtig als het gaat om het ontsluiten van gestructureerde gegevensbronnen. Ongestructureerde gegevensbronnen vormen een ander verhaal en zullen meer hoofdbrekens opleveren. Hoe nu deze ongestructureerde gegevens toch gestructureerd te ontsluiten? Het laatste wat we willen is een ETL-tool ongestructureerd inzetten. Uiteraard zou voor iedere aanleverende partij een specifieke extractie (load) mapping gerealiseerd kunnen worden, maar als het aantal aanleverende partijen groot is, levert dit een lastig onderhoudbare situatie op. Een veel betere oplossing is het ontleden van de ongestructureerde input en de technische metadata op te slaan in een



Afbeelding 3: Gemarkeerde data-elementen worden gemapped op de gewenste structuur.

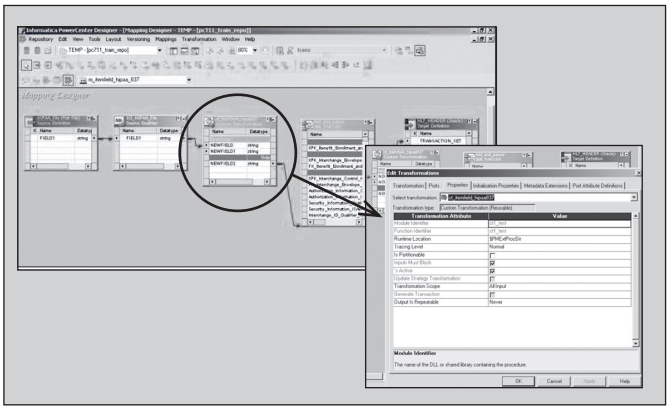
gestructureerd configuratiebestand. Zo kan voor iedere input worden vastgelegd welke – en op welke wijze – gegevens worden aangeleverd. De in deze paragraaf beschreven oplossing noemen we de configuratiesheet-oplossing. In afbeelding 1 is te zien hoe de ongestructureerde input van een bepaalde aanleverende partij wordt vastgelegd in een configuratiesheet. Dit kan overigens ook met behulp van een configuratie-databasetabel of een configuratie XML-bestand.

Om te beginnen wordt een element-type SRC_CHECK bepaald. Deze is noodzakelijk om de aangeleverde file te identificeren en vast te stellen of de structuur (gestructureerd of ongestructureerd) in ieder geval identiek is aan eerder gemaakte afspraken. In afbeelding 1 staat dat cell D1 de waarde LAGER moet bevatten voor klant CC.DE.SAP.0002023154. Hierdoor weten we dus precies (en controleren we) dat we te maken hebben met een bekend en in te lezen bestand.
Voor deze klant staat tevens vermeld dat hij een Excel-sheet aanlevert (SRC_EXT is XLS), dat de eerste twee rijen van zijn bestand de header vormen (HDR_SIZE is 2) en dat de gebruikte datum notatie YYYY-MM is. Vervolgens wordt met behulp van element-type FIX_VALUE klantspecifieke informatie opgenomen welke niet in het bestand wordt aangeleverd, zoals bijvoorbeeld het gegeven dat het een Duitse klant betreft (FIX_VALUE, NSO, DE). Wel aangeleverde gegevens kunnen in zowel een SRC_CELL als in een SRC_COL worden aangeleverd. Zo wordt de maand (START_DATE) waarop de gegevens betrekking hebben, vermeld in SRC_CELL E1 (uiteraard in het formaat zoals reeds eerder gespecificeerd). De artikelcode (CTN), description en de sales en voorraad worden als details aangeleverd (SRC_COL).

Nu alle informatie rond het ongestructureerde bestand gestructureerd is beschreven in de centrale configuratiesheet, kan in Informatica PowerCenter worden volstaan met slechts één extractie-mapping. In deze mapping wordt de configuratiesheet als centrale lookup opgenomen om te bepalen hoe zowel de header als de details van een aangeleverd bestand gekoppeld moeten worden aan het gewenste doel-datamodel, zie afbeelding 2. Dit bespaart veel ontwikkel- en onderhoudtijd binnen Informatica PowerCenter. Deze oplossing streeft ernaar dat wijzigingen enkel en alleen in de centrale configuratiesheet doorgevoerd moeten worden en dat de core extractie-mapping ongemoeid blijft.
Een nieuwe externe gegevensbron kan nu snel worden opgenomen binnen deze oplossing. De gegevens dienen alleen te worden opgenomen in de configuratiesheet en de reeds ontwikkelde Informatica mappings/transformaties zullen de gegevens uit deze nieuwe bron op een gestandaardiseerde wijze verwerken.

Informatica PowerCenter option Unstructured Data

De configuratiesheet-oplossing leent zich voor met name Excel- en flat-files. Maar wat nu als externe gegevens in een Word-document of in een Adobe Acrobat Reader-bestand worden



Afbeelding 4: Een transformatie voor het voorbewerken van ongestructureerde gegevens wordt opgenomen in de Informatica-mapping.

aangeleverd? De configuratiesheet-oplossing is dan niet toepasbaar. Het de aanleverende partij vragen de gegevens alsnog als flatfile aan te leveren is niet altijd een optie. Informatica biedt sinds kort met de optie Unstructured Data de mogelijkheid ook ongestructureerde gegevens in zelfs lastige bestandsformaten te ontsluiten. Bestanden die niet in de vorm van een tabel met kolommen worden aangeleverd kunnen desondanks worden gemapped naar de gewenste structuur, zie afbeelding 3.
Met behulp van deze optie kan in een ongestructureerd bestand met behulp van search-types worden gezocht naar de gewenste gegevenselementen. Door als het ware het bestand met een marker te doorlopen kan de gezochte content worden geïdentificeerd en gemapped. Deze pretransformatie kan vervolgens worden opgenomen in een mapping zoals in afbeelding 4 duidelijk wordt.

Conclusie

Integreren van externe gegevens wordt steeds belangrijker. Externe gegevens worden vaak in diverse vormen en vaak ongestructureerd aangeleverd, aangezien het afdwingen van een standaard interface een lastige en niet altijd mogelijke optie blijkt. Problemen bij de verwerking kunnen zijn: de kwaliteit van de gegevens; de betekenis van de gegevens; de relatie leggen tussen externe en interne gegevens. Bovenstaande problemen kunnen in meer of mindere mate worden voorkomen door te streven naar een win-win situatie voor zowel de vragende als aanbiedende partij van externe gegevens.
De configuratiesheet-methode laat zien hoe het mogelijk wordt (externe) niet gestructureerde gegevens, aangeleverd in bepaalde formaten, gestructureerd te beschrijven om ze vervolgens te kunnen ontsluiten en te kunnen integreren met interne gegevens. Tot slot wordt gewezen op het feit dat er nieuwe technische hulpmiddelen beschikbaar zijn voor het ontsluiten en integreren van (externe) niet gestructureerde gegevens, voor een groot aantal verschillende formaten, waardoor het automatisch verwerken van externe niet-gestructureerde gegevens mogelijk wordt.

Sjoerd Hobo (sjoerd.hobo@qnh.nl) en **Rob Peters** (rob.peters@qnh.nl) zijn senior consultant bij QNH Enterprise Intelligence bv.