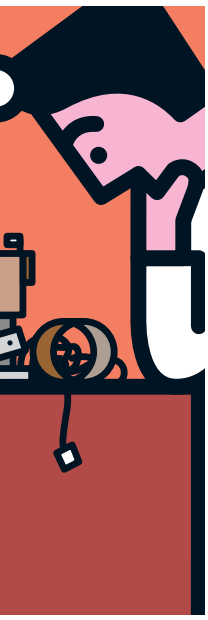


Een duidelijke visie voor Enterprise Data Integration is nodig

Op zoek naar controle

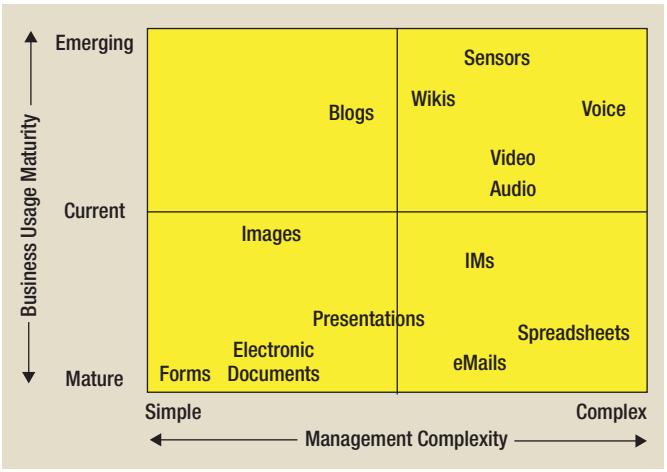
Erwin Vorwerk

Meer dan 80 procent van alle beschikbare data valt volgens analisten in de categorie ongestructureerde data. Pas sinds een aantal jaren is er meer aandacht voor ongestructureerde data. Maar is dat niet te laat?

De chaos rondom ongestructureerde data in organisaties neemt sterk toe, een toenemend gedeelte van de data wordt ongecontroleerd opgeslagen (bijvoorbeeld mensen gebruiken in toenemende mate hun inbox als informatiedatabases), mensen besteden meer tijd aan het zoeken naar data dan het analyseren ervan. Waarom zijn ongestructureerde data belangrijk en wat kunnen we ermee? Hebben we de controle niet al lang geleden verloren?

Ongestructureerde data zijn data die gecreëerd en opgeslagen worden zonder gebruik te maken van formele templates of formulieren; zowel analoog als digitaal. We beperken ons in dit artikel vanzelfsprekend tot de digitale variant, die een aantal verschijningsvormen kent, zie afbeelding 1.

Als een organisatie succesvol wil zijn, dan zullen zowel de gestructureerde data als de ongestructureerde data op een professionele en robuuste manier moeten worden beheerd en beheerst. Nog lang niet alle organisaties hebben dit stadium bereikt, maar steeds meer directies realiseren zich dat hun Business Intelligence zich beperkt tot slechts 20 procent van de binnen een onderneming aanwezige data en dat men dus op basis van een beperkte blik beslissingen neemt. Daarmee neemt de vraag naar geïntegreerde informatie toe.



Afbeelding 1: Ongestructureerde data afgezet tegen volwassenheid en complexiteit.

Optimaal

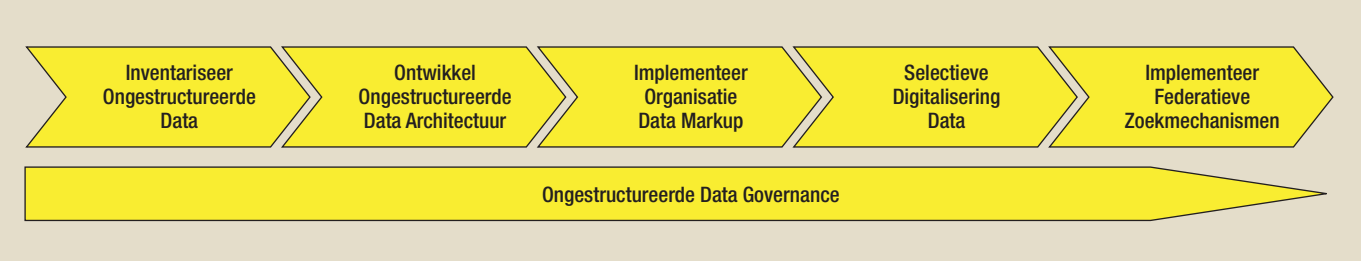
Het bereiken van het optimaal beheren en beheersen van ongestructureerde data kan worden gezien als een project, uitgedrukt in een aantal fases. Zie afbeelding 2.

Data governance van ongestructureerde data regelt vanuit de organisatorische strategie de processen, kennis en kunde, leiderschap en middelen die nodig zijn om de informatiebronnen van een organisatie succesvol en structureel te beheren en beheersen. Data governance voor ongestructureerde data moet net zo rigide zijn als data governance voor gestructureerde data en moet integraal onderdeel gaan uitmaken van de bedrijfsbrede data governance.

De belangrijkste initiële taken voor governance van ongestructureerde data zijn:

1. het opnemen van ongestructureerde data in de bedrijfsbrede data governance organisatie en uitvoering;
2. het starten van een groep die overzicht heeft over zoekmechanismen en rapportages;
3. het starten van een audit-functie om de kwaliteit van ongestructureerde data te maximaliseren.

Als de data governance voor ongestructureerde data onvoldoende aandacht krijgt, dan zal het initiële succes van het project snel verbleken doordat de waarde van ongestructureerde data afneemt, vanwege gebrek aan inbedding van belangrijke processen (bijvoorbeeld kwaliteitscontroles, eigendom) in de organisatie.



Afbeelding 2: Fasering van project ter beheer en beheersing van ongestructureerde data.

Opslag en toegang

De hamvraag is waar(mee) en hoe ongestructureerde data op te slaan. Kan dit in de bekende relationele databases, of op file servers, of content management-systemen, of kan het ook met een combinatie van deze manieren? Voor het beantwoorden van deze vraag moet rekening worden gehouden met een aantal belangrijke facetten:

- wet- en regelgeving ten aanzien van bewaartermijn, traceerbaarheid, etcetera;
- identificatie en classificatie van de inhoud;
- toegankelijkheid en gebruik van de data;
- integratie van ongestructureerde en gestructureerde data;
- visie en beleid op het vlak van informatie management.

Er zijn twee technologische hoofdstromingen waarin ongestructureerde data kunnen worden opgeslagen: Enterprise Content Management systemen en relationele databases. Hieronder volgt een korte beschrijving en analyse van beide stromingen.

Enterprise Content Management

Voor het werken met ongestructureerde data wordt vaak gebruik gemaakt van Enterprise Content Management-oplossingen (ECM), een paraplubegrip voor een aantal aparte, maar gerelateerde technologieën zoals *document management, web content management, digital asset management, records management* en onderdelen van *enterprise search* en *collaboration*.

> Ook in relationele databases worden steeds vaker ongestructureerde data opgeslagen

Veel gebruikte oplossingen voor het beheren en beheersen van ongestructureerde data zijn de zogenaamde document management oplossingen. Deze oplossingen bieden vooral mogelijkheden op het vlak van het opslaan van de historie van veranderingen aan een document, en het in- en uitchecken van documenten door gebruikers om versieconflicten te voorkomen en de autorisatie voor toegang/wijziging van documenten.
Vaak bestaan er ook uitgebreide mogelijkheden om additionele metadata over documenten in te voeren, zoals auteur, sleutelwoorden, datums, project, informatiedeelgebied, etcetera.

Meer in de richting van digital asset management en record management worden ook mogelijkheden geboden als het beheersen van workflow en de levenscyclus van documenten, wat met name van belang is in het kader van het naleven van regelgeving zoals Sarbanes-Oxley. Zie afbeelding 3.

Relationele databases

Ook in relationele databases (datawarehouses) worden steeds vaker ongestructureerde data opgeslagen. Veelal bieden relationele databases de mogelijkheid om ongestructureerde data (of verwijzingen naar ongestructureerde data) op te slaan in de vorm van een BLOB (binary large object). Maar dit is onvoldoende om zinvol met ongestructureerde data om te gaan, omdat het slechts om een soort container gaat waarin deze data kunnen worden opgeslagen.
Om ongestructureerde data in relationele databases zinvol te kunnen gebruiken, is het noodzakelijk (net zoals in ECM-oplossingen) dat er metadata aan deze objecten worden gerelateerd. Deze metadata moeten beschrijven wat de inhoud van de desbetreffende ongestructureerde data is (sleutelwoorden die de inhoud van het document aan een specifiek thema koppelen) en wat de context van de ongestructureerde data is (informatie over status, datum, auteur, etcetera).

Voordelen	Nadelen
- Goede performance ongestructureerde data - Geavanceerde metadata, tagging, etc. - Eenvoudige integratie van grote verscheidenheid aan data	- Specifiek archiverings/backup mechanisme noodzakelijk - Replicatie kan erg lastig zijn - Integratie met gestructureerde data niet altijd volwassen

Afbeelding 3: Voordelen/nadelen opslag ongestructureerde data in Enterprise Content Management-systemen.

Voordelen	Nadelen
- Alle data staan op één plaats - Schaalbaarheid, clustering - Eenvoudige replicatie - Archivering/backup ingebouwd	- Niet altijd geoptimaliseerd voor het werken met ongestructureerde data - Explosie databasevolume

Afbeelding 4: Voordelen/nadelen opslag ongestructureerde data in relationele databases.

Karakteristiek	ETL	EII	EAI
Werkwijze	Batch	Real-time	Real-time
Aansturing	Vast schema	Query	Gebeurtenis
Verplaatsing data (1)	Vóór de query	Tijdens query	Vóór de query
Verplaatsing data (2)	Vaste set data	Alleen benodigde data	Vaste set data (berichten)
Verplaatsing data (3)	Fysieke verplaatsing naar andere opslag (bv. DWH, ODS, datamart)	Geen fysieke dataverplaatsing, alleen doorgifte aan presentatielaag	Fysieke verplaatsing naar andere opslag (bv. applicatie-database)
Mechanisme	Pull	Pull	Push
Toepassing	Hoofdzakelijk rapportage en analyse	Hoofdzakelijk rapportage en analyse	Hoofdzakelijk *business process automation*
Datavolumes	Geschikt voor grote volumes	Geschikt voor kleine datavolumes (krachtig bij kleine ad hoc query's)	Geschikt voor klein/middelgroot datavolume
Prijsontwikkeling & onderhoud	Normaal gesproken niet zo duur als EAI	Over het algemeen goedkoper dan ETL en EAI (erg geschikt voor iteratieve ontwikkeling)	Normaal gesproken de duurste optie voor integratie

Afbeelding 5: Vergelijking integratie-technologieën.

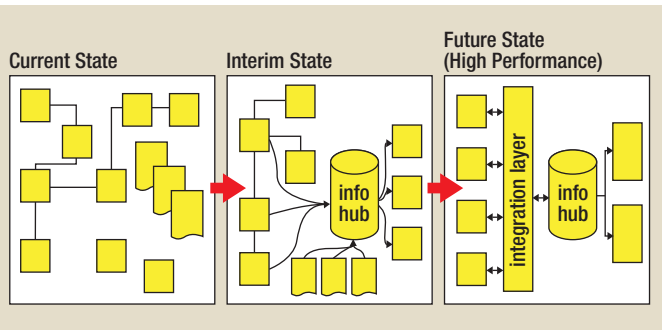
In de praktijk wordt veel gewerkt met twee verschillende datawarehouses. Het traditionele (gestructureerde) datawarehouse bevat alleen de belangrijkste metadata van de ongestructureerde data, waardoor integratie van informatie vanuit beide omgevingen mogelijk wordt. Het ongestructureerde datawarehouse bevat gedetailleerde metadata en (koppelingen naar) de feitelijke metadata. Veelal wordt – door het analyseren van het gebruik en toegevoegde waarde van ongestructureerde data – bepaald of de ongestructureerde data feitelijk in de database worden opgeslagen of dat een link wordt opgeslagen naar de fysieke of logische locatie van het document. Zie afbeelding 4.

De strategisch juiste keuze

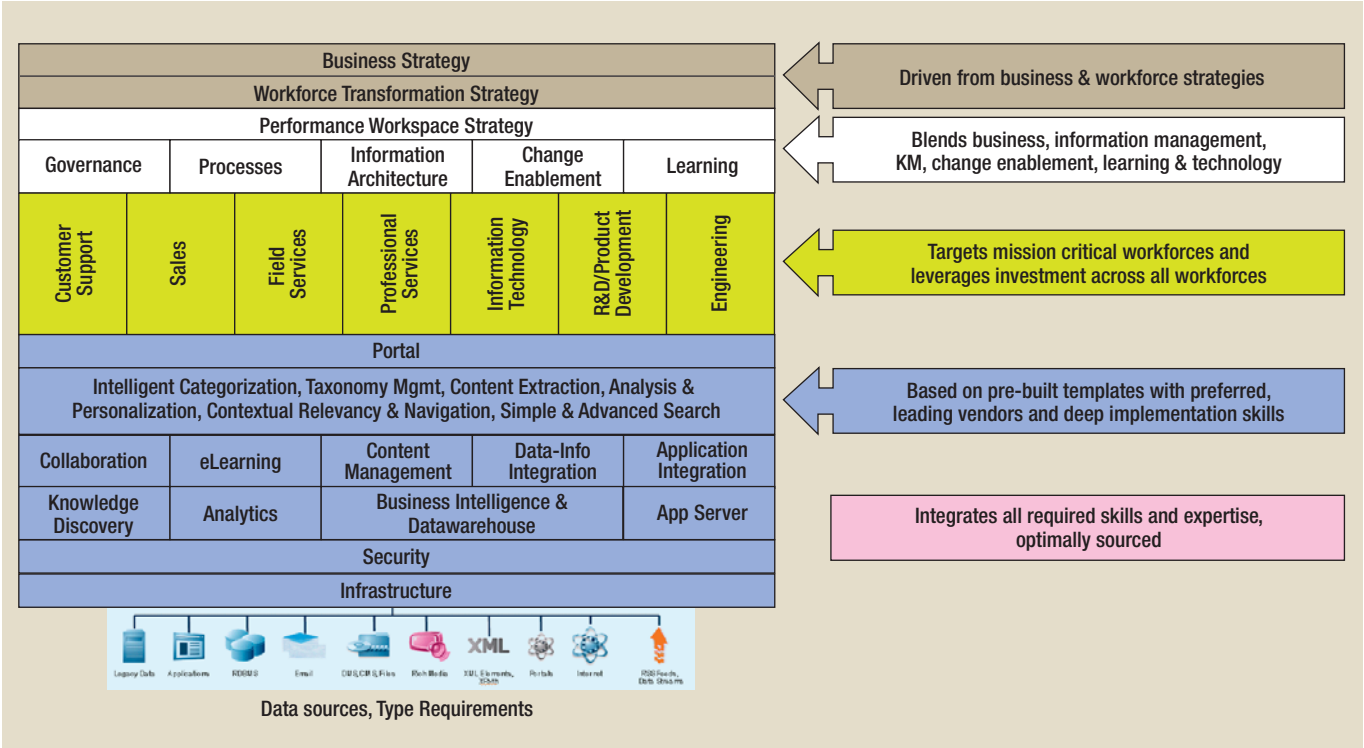
Ongeacht de technologie die voor het werken met (on)gestructureerde data toegepast wordt, zal men vanuit een duidelijke visie *Enterprise Data Integration* moeten toepassen om het succesvol te maken. Hierbij zullen de volgende stappen moeten worden doorlopen:

- Het identificeren en rangschikken van processen en applicaties naar toegevoegde waarde voor de business;
- Het in kaart brengen van databronnen die van groot/kritiek belang zijn;
- Het in kaart brengen, begrijpen en ontwikkelen van metadata;
- Het ontwikkelen van een datamodel dat gestructureerde en ongestructureerde data met elkaar in verband kan brengen;
- Het bieden van structuur en analysemogelijkheden voor ongestructureerde data;
- Het maken van een synthese tussen automatische en handmatige verwerking van media;
- Het werken met software-leveranciers die in staat zijn de visie op unificatie uit te voeren;
- Het implementeren van een losse koppeling tussen inhoud en applicaties;
- Het aansturen van al deze activiteiten via de principes en strategie van een *Enterprise Information Management* programma.

De keuze met betrekking tot de technologie voor opslag van informatie zou niet het vertrekpunt in de discussie over de integratie tussen ongestructureerde en gestructureerde data moeten zijn. In afbeelding 5 staan de belangrijkste verschillen tussen de meest gebruikte technologieën voor informatie-integratie op een rij. Een synthese van de verschillende technologieën behoort vaak ook tot de mogelijkheid, omdat het in de eerder geschetste informatiearchitectuur in principe niet uitmaakt waar de informatie vandaan komt.



Afbeelding 6: Enterprise Information Integration.



Afbeelding 7: Mogelijke informatie management architectuur.

De ideale oplossing ligt niet zozeer in het onderbrengen van alle data, of het nu gestructureerde of ongestructureerde data zijn, in één technologie of bron. Veel betere resultaten kunnen worden verwacht van het integreren van verschillende informatiebronnen op basis van consistente metadata via een informatie-hub. Zie afbeelding 6.

Via deze benadering wordt de informatie aangeboden via een portal-achtige oplossing en is het voor de gebruikers transparant waar de benodigde informatie vandaan komt en kan via een integratielaag worden gezorgd voor deze transparantie. De informatie-hub bevat feitelijk een catalogus van de beschikbare informatie en samenvattingen van de meest gebruikte informatie.

> Er zijn grofweg drie categorieën van analyses die op ongestructureerde data kunnen worden uitgevoerd

De informatie wordt opgeslagen in het meest geschikte systeem en door het loskoppelen van data en applicatie kan vrij eenvoudig worden gemigreerd naar meer geschikte technologieën.

Maar als we dit Enterprise Information Integration concept goed willen toepassen, dan moet dit over de hele linie van informatiesystemen gebeuren. Het vertrekpunt voor het realiseren van dergelijke trajecten is de bedrijfsstrategie en de transformaties die men via deze strategie wil bereiken. Of het nu om ongestructureerde of gestructureerde data gaat, alle benodigde data worden aan specifieke doelgroepen aangeboden voorzien van de benodigde functionaliteit.

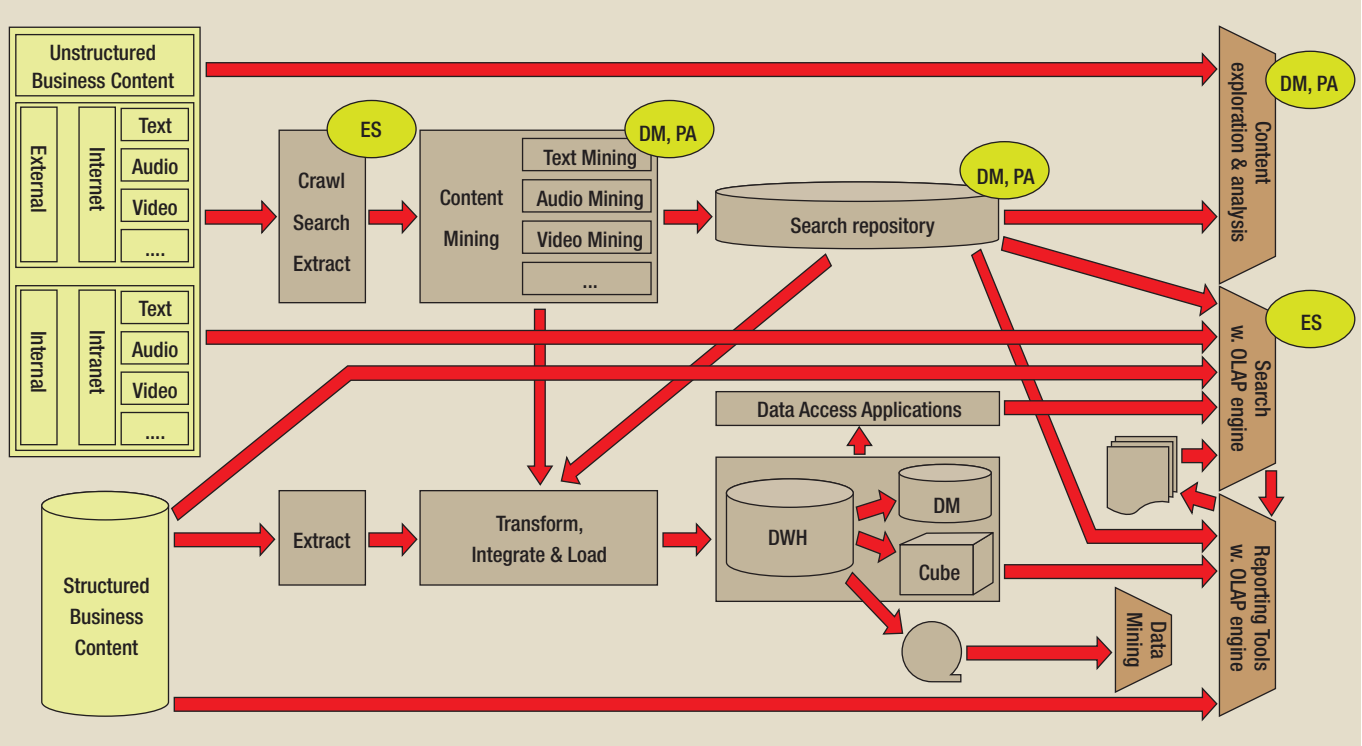
Zoals uit afbeelding 7 blijkt, ligt het zwaartepunt op het bieden van de juiste informatie aan de juiste mensen en zijn de feitelijke data van secundair belang: deze zijn immers via de beschikbare infrastructuur probleemloos te benaderen.

Realisatie

In de (ideale) informatie management architectuur van afbeelding 7 worden de gestructureerde en ongestructureerde informatie via de onderkant van de infrastructuur met elkaar in verband gebracht door middel van een *data integration engine*. Deze engine maakt primair gebruik van open standaarden om data uit te wisselen, zoals webservices.

Deze engine bestaat uit een verscheidenheid aan geïntegreerde componenten (er zijn immers nog geen componenten die alle taken op dit vlak volledig kunnen uitvoeren). Deze componenten zorgen ervoor dat de data geëxtraheerd en getransformeerd worden en dat de relaties en de context van de data geanalyseerd worden. Daarnaast worden taken door de engine uitgevoerd op het vlak van datakwaliteit en het verrijken van data. Idealiter heeft deze engine een mechanisme voor gedistribueerde verzameling en distributie van data.

De portal fungeert als interface voor de bezorging van data aan de verschillende gebruikersgroepen die op basis van hun rol toegang krijgen tot de data, via een door de gebruiker zelf te kiezen medium (browser, e-mail, PDA, SMS). De data vinden de gebruiker in plaats van andersom. Op basis van de momenteel bestaande technologie is een dergelijke oplossing nog steeds



Afbeelding 8: Analyse van ongestructureerde data gepositioneerd in het informatie-analysemodel.

een samensmelting van verschillende tools. Met de recente overnamebewegingen in de markt zijn enkele leveranciers (Oracle, IBM, Microsoft) wel dichter bij het doel gekomen; het ultieme doel van één oplossing waarin gestructureerde en ongestructureerde data op één consistente en transparante manier in één omgeving beheerd en gebruikt kunnen worden. Analyse van data is een onderwerp dat de laatste tijd zeer sterk in de belangstelling staat. In de markt zijn vele tools beschikbaar voor het analyseren van data. Echter, niet alle tools kunnen adequaat omgaan met het analyseren van ongestructureerde data. Er zijn grofweg drie categorieën van analyses die op ongestructureerde data kunnen worden uitgevoerd: *data mining, predictive analysis* en *enterprise search*. Het voert te ver om deze categorieën in het kader van dit artikel verder uit te werken.

Als we diverse analysemogelijkheden ten aanzien van ongestructureerde data projecteren op het informatie-analysemodel, dan ontstaat het beeld zoals in afbeelding 8. Hieruit blijkt dat Business Intelligence voor een organisatie pas compleet is wanneer de ongestructureerde data integraal onderdeel uitmaken van het informatie-aanbod ter ondersteuning van de besluitvorming.

De toekomst

Het beheren en beheersen van ongestructureerde data zal zich blijven ontwikkelen. In vele opzichten staat het nog in de kinderschoenen en heeft het naar schatting een achterstand van twintig jaar op het beheren en beheersen van gestructureerde data. Een aantal trends op een rij.

Een holistische benadering is van groot belang.
Er moet een juiste balans gevonden worden tussen organisatie, governance, processen en architectuur. Het structureel integreren van ongestructureerde data in bestaande processen en applicaties biedt de meeste toegevoegde waarde. Een aantrekkelijke terugverdientijd van de investeringen is snel te bepalen.

Het is werk in uitvoering, in het stadium van early adopters.
Er is op dit moment geen enkel product dat alle informatie management-behoeften kan dekken. Hoewel de benodigde technologie beschikbaar is, blijft het gebruiken van verschillende producten op dit moment noodzakelijk. Pas bij de volgende generatie tools komt het naadloos integreren van gestructureerde en ongestructureerde data in zicht.

System integrators moeten geïntegreerde infrastructuur leveren.
Voor het adequaat samenbrengen van gestructureerde en ongestructureerde informatie is een samensmelting nodig van diepgaande kennis, methodologieën, tools en samenwerkingsverbanden. Omdat organisaties hier vaak niet in willen en kunnen investeren, is hier voor system integrators een belangrijke rol weggelegd.
Maar voordat Enterprise Information Integration-concepten zinvol kunnen worden toegepast is het noodzakelijk dat de ongestructureerde data goed beheerd en beheerst worden en in verband kunnen worden gebracht met gestructureerde data.
Een klus die niet eenvoudig zal zijn, maar wel belangrijke toegevoegde waarde biedt voor het succes van de organisatie.

Erwin Vorwerk (erwin.vorwerk@accenture.com) is Senior Manager bij Accenture.