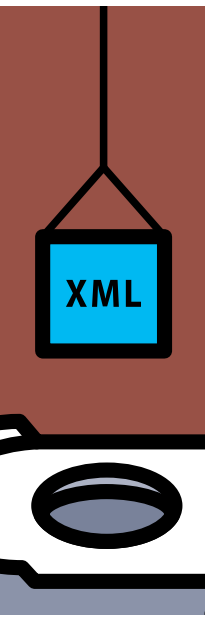


Jan-Willem de Koning van SGML/XML User Group Holland:

XML is zó succesvol dat het aan het verdwijnen is

Robbert Hoeffnagel

"We gaan van techniek naar semantiek", zegt Jan-Willem de Koning. Hij is voorzitter van de SGML/XML User Group Holland en in het dagelijks leven actief als manager content operations bij SDU Information Services in Den Haag.

"Ik hoor aanbieders van search engines wel eens voorspellen dat XML de relationele database overbodig gaat maken. Dat lijkt me onzin. Maar als ik kijk naar wat er bijvoorbeeld met DB2 gebeurt of op het gebied van RSS, dan zie ik wel dat XML een steeds grotere rol speelt als het om het ontsluiten van informatie gaat."
Een gesprek met Jan-Willem de Koning van de SGML/XML User Group Holland over 'content die onderin zit', DTD's en schema's, en de steeds grotere rol die semantiek speelt.

> Een DTD is een informatie-model dat je ook in een database kwijt kunt

"Op technisch gebied mogen we XML zo langzamerhand wel volwassen noemen. Dat geldt echter nog zeker niet voor de vele standaarden – of pogingen daartoe – die op XML zijn gebaseerd. Dat zie ik ook als ik naar de activiteiten van de Nederlandse gebruikersgroep kijk. Vroeger ging het eerst en vooral om de techniek en was het soms meer zendingswerk om mensen uit te leggen wat het is en wat je er mee kunt. De kreet 'XML' was in die tijd bij wijze van spreken al genoeg om een volle zaal te trekken. Nu gaat het eigenlijk alleen nog maar om de toepassingen en dan liefst ook nog toegespitst op het gebruik van XML in een specifieke branche. Het gaat dus meer en meer om de semantiek."

Afgeleid van SGML

De Koning is werkzaam bij de SDU, waar hij zich onder andere bezig houdt met wat wel 'de officiële publicaties' van de overheid worden genoemd. Denk aan de staatscourant en dergelijke, maar ook een website als wetten.nl valt onder zijn verantwoordelijkheid. In het gesprek hanteert De Koning regelmatig termen die de historie van XML aangeven. XML is voortgekomen uit SGML – de Standard Generalized Markup Language – een metataal voor het definiëren van markup languages. Hij corrigeert zichzelf regelmatig als hij in plaats van bijvoorbeeld 'schema' het uit SGML afkomstige begrip DTD ofwel Document Type Definition gebruikt. "Dat is natuurlijk hetzelfde, alleen gebruikt iedereen tegenwoordig de term schema, die in de ICT-wereld ook veel gebruikelijker is."
Toch is het wel een treffende 'verspreking'. Het geeft namelijk goed aan dat XML voor veel mensen nog altijd hele andere dingen betekent.

"XML kent inmiddels drie gebruiksdoelen", legt De Koning uit. "Het eerste is het opslaan van content. We kunnen hiermee betekenis geven aan elementen in die content. Dit zien we veel bij bijvoorbeeld uitgeverijen gebeuren. Het tweede toepassingsgebied is in de wereld van de ICT het bekendst: de rol van uitwisselingsformaat. Maar, zeg ik er direct bij, wat ons betreft is XML dan 'just another messaging format'. XML is immers niet eens zo'n heel bijzonder formaat. Met Comma Separated Values komen we ook een heel eind als het op uitwisselen aankomt. Alleen is bij gebruik van XML de content gewoon leesbaar, terwijl we bij CSV-bestanden wat meer het risico lopen dat de inhoud van het bestand in eerste instantie minder begrijpelijk is. Een lijst van woorden of getallen zegt natuurlijk vrij weinig, tenzij je de betekenis daarvan kent. Een XML bestand omvat zowel de content als de betekenis van die content. Dat is ook erg handig als werk bijvoorbeeld aan een outsourcing-partner in een ander land wordt overgedragen. De content is voor die partij dan vrijwel direct te begrijpen.

Het derde gebied waarin XML een belangrijke rol speelt, is bij de communicatie tussen applicaties. In een SOA/ESB (Service Oriented Architecture/Enterprise Service Bus) omgeving dus. Van de drie toepassingsgebieden zien we dit nog het minst gebeuren."

Gartner's hype cycle

Dat zijn op het eerste gezicht drie zeer verschillende toepassingsgebieden. Wat heeft het vastleggen van content als uitgeefprobleem immers te maken met XML gateways die twee of meer legacy-systemen via een centrale bus met elkaar laat 'praten'? Of het gebruik van XML als vervanger voor de aloude Edifact-berichten. Toch klopt die perceptie niet, meent De Koning. "In alle gevallen gaat het wel degelijk om hetzelfde. XML is een taal waarmee content op een goed gestructureerde manier kan worden beschreven, waardoor het voorspelbaar en dus automatiseerbaar wordt. Veel meer dan dat is XML niet."
In eerste instantie lijkt XML wellicht complex en abstract, maar in de praktijk is het dat niet. "We hebben het over een methode waarmee we elementen in content kunnen aangeven en definiëren. Die content kan een document zijn dat een uitgever publiceert, maar ook een bericht dat tussen twee informatiesystemen wordt uitgewisseld. Het is in feite alleen maar een afspraak om bepaalde typen informatie te definiëren. Juist het feit dat we structuur aan content kunnen meegeven, maakt XML overigens zo krachtig."

Tegelijkertijd worden met grote regelmaat op XML gebaseerde standaarden of voorstellen daartoe gelanceerd. Die maken dat we soms door de bomen het bos niet meer zien. Wie Gartner's fameuze 'hype cycle' van enige jaren geleden op het gebied van XML ziet (afbeelding 1), zal het direct opvallen. SAML, XSLT, XBRL, UBL, ebXML, WSDL – het is een lange rij van afkortingen die alle een andere 'markup language' vertegenwoordigen.
"In feite zijn het niet meer dan afspraken voor coderingsschema's: welke betekenis moeten we aan een bepaald element in content toekennen? Anders gezegd: hoe geven we aan wat een bepaald element in de content nu precies betekent? En dat wordt dan iedere keer toegespitst op een specifiek doel of een bepaalde branche of toepassingsgebied. Waar XML als – zeg maar – onderliggende technologie inmiddels volwassen is, zijn al deze standaarden dat echter vaak nog veel minder. Al verdienen SOAP, XSLT, XQuery, ebXML/XBRL en bijvoorbeeld XSL-FO die kwalificatie inmiddels wel."

Taxonomieën en ontologieën

XML speelt een steeds grotere rol bij het ontsluiten van data en informatie. "Mensen die zijn opgegroeid met databases lijken vaak aan het idee te moeten wennen, maar hooguit twintig procent van alle informatie kan en wordt in databases vastgelegd. We leggen alleen maar die data in een database vast die daar ook geschikt voor zijn. Voor die andere tachtig procent geldt dat het ongestructureerde informatie is waarvoor we iets

Foto: Harry Otto.

Jan-Willem de Koning: "Het structurerend vermogen zit in de search engine en niet in de metadata die aan documenten wordt meegegeven".

anders moesten verzinnen. Juist daar zit de toekomst van XML. Eigenlijk zou XML mensen uit de wereld van de databases erg moeten aanspreken en eerlijk gezegd zie ik dat ook wel gebeuren. XML structureert immers content, die daardoor makkelijk te ontsluiten is."
Tegelijkertijd zijn in de markt nogal tegenstrijdige geluiden te horen over de relatie tussen XML en databases. "Aan de ene kant zien we de zoekmachines. En nee, dan bedoel ik niet Google", zegt De Koning. "Voor informatie-professionals is Google eerlijk gezegd weinig meer dan een redelijk eenvoudige zoekmachine die vooral snel is. De meest interessante vooruitgang op het gebied van search vindt elders plaats. Ik denk dat de ontwikkelingen die we zien bij Autonomy, Fast, het Nederlandse Irion of bijvoorbeeld Endeca veel relevanter zijn. Wat je daar ziet gebeuren, is dat er – gechargeerd gezegd – helemaal niet meer gestructureerd wordt. In plaats van tags en metadata toe te voegen, richt de aandacht zich daar veel meer op de taxonomie, de ontologie of wat voor kreet men er ook voor heeft verzonnen. Met andere woorden: de woordenlijsten die dit

soort business search engines gebruiken. Het structurerende vermogen zit in de search engine en niet in de metadata die aan documenten wordt meegegeven."

De Koning is onder andere verantwoordelijk voor een uit acht medewerkers bestaande redactie die juridische informatie van metadata voorziet. "Eigenlijk zou ik liever zien dat zij zich full-time konden bezig houden met het onderhouden van taxonomieën. Op die manier kan informatie immers heel gericht worden ontsloten en zouden we bijvoorbeeld veel meer portals kunnen inrichten die zich specifiek op een bepaalde doelgroep richten. Alles gebaseerd op dezelfde juridische content die iedere keer met een net iets andere taxonomie wordt doorzocht. Iedere keer dus een andere 'view' op dezelfde content. Soms denk ik wel eens dat informatie ontsluiten via search net een exotische vorm van Business Intelligence is."

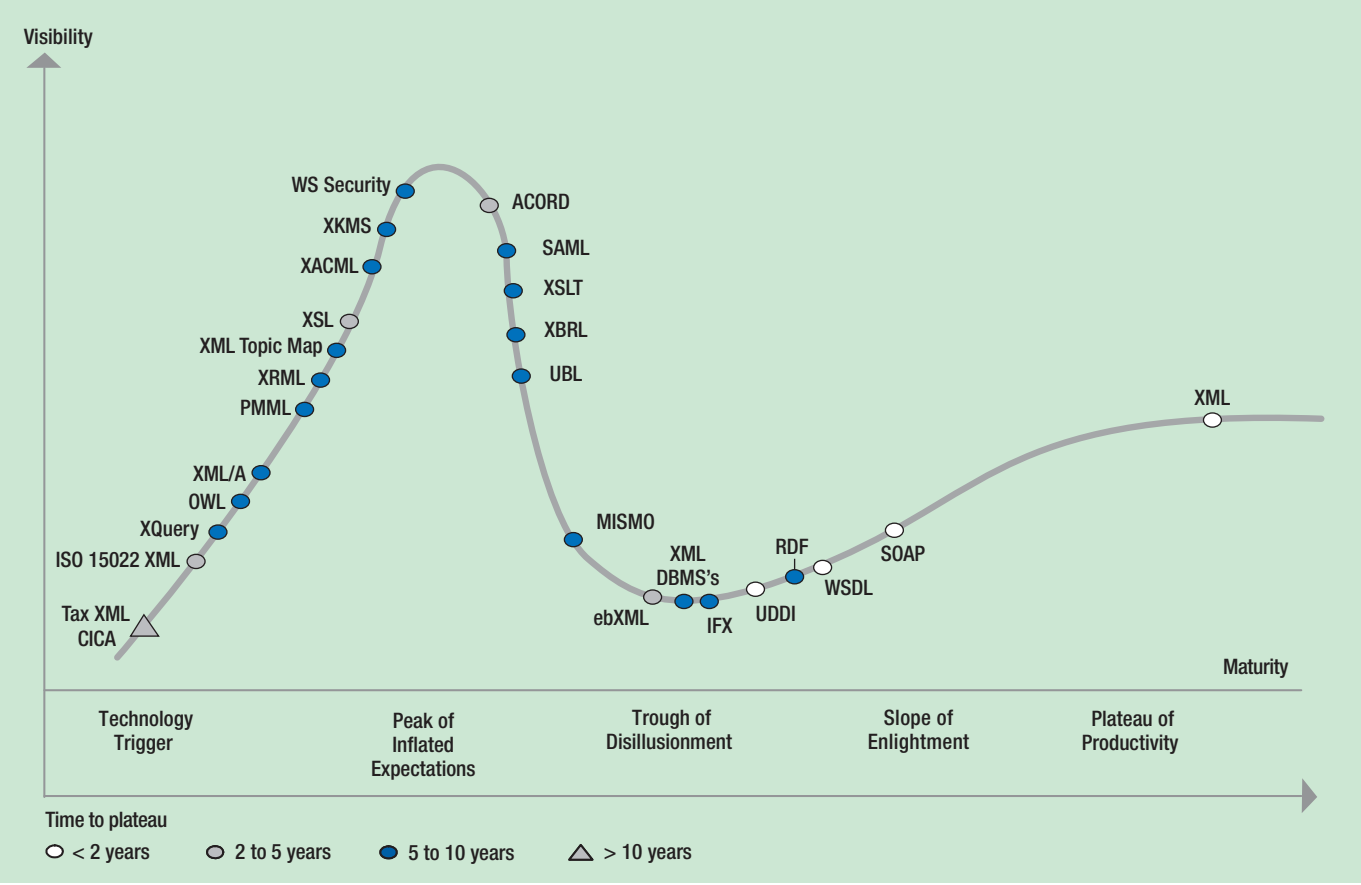
Tegenstelling

"Ik sprak enige tijd terug mensen van Fast. Dat is de opvolger van het bekende AltaVista, maar is nu geheel op business-toepassingen gericht. In die gesprekken werd de voorspelling gedaan dat we over vijf jaar helemaal geen databases meer nodig hebben. Search neemt alles over, zeggen zij. Dat lijkt me een vooral door marketing ingegeven uitspraak, maar het is natuurlijk wel waar dat de ontwikkelingen op het gebied van zoektechnologie erg hard gaan en zeer interessant zijn voor iedereen die zich met het ontsluiten van zowel gestructureerde als ongestructureerde informatie bezig houdt."

> In plaats van tags toevoegen richt de aandacht zich veel meer op taxonomie

Aan de andere kant van het spectrum zien we bedrijven als IBM en Oracle die XML 'native' in hun database willen stoppen. De laatste onder andere via het overgenomen SleepyCat. Met name de hybride producten die beide omgevingen aan kunnen, lijken daarbij hoge ogen te gaan gooien. Daarnaast bestaat een aantal 'pure players'. Een voorbeeld is de XML database die Mark Logic levert. Die zijn met name interessant voor die eerste groep van toepassingen die De Koning noemde en zullen hun weg voornamelijk naar uitgeverijen vinden.

"Er bestaat dus een tegenstelling tussen enerzijds een stroming die volledig wil structureren en anderzijds een groep die eigenlijk helemaal niets meer aan structuur wil aanbrengen.



Afbeelding 1: Gartner's hype cycle voor XML en op XML gebaseerde standaarden.

Mijn persoonlijke mening is dat we het hier niet over een of/of situatie hebben maar over en/en. Beide omgevingen zullen naast elkaar blijven bestaan, maar hier en daar ook in elkaar opgaan. Welke technologie het beste is voor het ontsluiten van informatie, valt niet te zeggen. Zij dienen verschillende doelen. Bovendien is eerder sprake van haasje-over waarbij momenteel de searchjongens dan misschien de bovenliggende partij lijken te zijn, maar dat kan volgend jaar weer heel anders zijn."

Hybride oplossingen

"Neem wetgeving en juridische informatie. Hoe graag ik mijn mensen ook op het beheren van taxonomieën zou zetten, er is geen jurist die het prettig vindt om te werken met langs geautomatiseerde weg gegenereerde content. Hij wil honderd procent zeker weten dat de jurisprudentie die hem ten aanzien van een bepaald wetsartikel wordt aangereikt ook daadwerkelijk klopt, de link correct is en de informatie relevant. Dat kan dus alleen als de informatie door een juridisch geschoolde redacteur is voorbereid, van metadata is voorzien en – zeg maar – is klaargezet. Structuur is hier dus van cruciaal belang. Aan de andere kant staan bedrijven als Reuters die aan de hand van profielen van hun abonnees binnenkomende persberichten volledig geautomatiseerd doorsturen. Daar komt geen mens aan te pas. En daartussen zit een groot grijs gebied." XML gaat de rol van de database dus absoluut niet overnemen, meent De Koning. Maar in hybride oplossingen ziet hij wel veel. "Met XQuery kun je natuurlijk prima beide werelden aan elkaar knopen. In een database mogen dan kale data liggen die pas waarde of betekenis krijgen nadat deze in een applicatie zijn ingebracht, via een XML markup kun je uitstekend betekenis aan die kale data geven. Je kunt dus wel degelijk met XML data in een database benaderen. Dat kan in allerlei situaties handig zijn, maar wat ik vooralsnog niet zie gebeuren is het bouwen van krachtige applicaties op basis van XML als ontsluitingsmechanisme. Ik beschouw zo'n uitspraak over de verdringing van de database door XML dan ook maar vooral als een signaal dat er tal van bewegingen in de markt gaande zijn."

Beide omgevingen zullen hier en daar in elkaar opgaan, stelt De Koning. "Een DTD is immers weinig anders dan een informatiemodel dat je ook in een database kwijt kunt. Het probleem is alleen dat de content dan als het ware 'onderin zit' en je dus eerst pakweg twintig tabellen door zult moeten voordat je bij de data bent die je nodig hebt. Dat maakt dat het ook zo traag is en vandaar natuurlijk het gebruik van Blob's. Een bedrijf als X-Hive – onlangs overgenomen door EMC, dat daarmee toch een duidelijk signaal gaf dat XML het predicaat 'volwassen' verdient, zie ook pagina 24 in dit blad – heeft echter laten zien dat je wel degelijk goed kunt filteren en dergelijke. Je kunt met XML in feite van alles 'aan' of 'uit' zetten, om het maar zo uit te drukken. Dat is immers een kwestie van tags toevoegen."

Slechte tools

XML heette lange tijd complex, traag en duur te zijn. "Dat is natuurlijk helemaal niet waar en zegt eerlijk gezegd meer over de historie dan over de techniek zelf. Die reputatie is ontstaan toen uitgeverijen met DTD's en dergelijke aan de slag wilden gaan. Uitgevers zagen in XML een mooi middel om content op een medium-neutrale wijze vast te leggen, om vervolgens dezelfde informatie via tal van kanalen te publiceren. Tegelijkertijd beschikten zij over veel bestaande content. Al die bestaande artikelen moesten eerst aan de hand van DTD's worden gestructureerd. Dat is natuurlijk een enorme operatie die veel tijd en geld kost. Daar komt die reputatie van traag en duur vandaan. Met de feitelijke techniek heeft dat niets te maken. XML is natuurlijk ook weinig meer dan een plat ascii-bestandje. XML verwerken gaat juist heel snel en het toevoegen van tags – mits het in het normale werkproces van de gebruikers is opgenomen – kost nauwelijks tijd."

Toen XML opkwam ontbrak het echter simpelweg aan tools. "En de paar tools die er waren, waren gewoon slecht. Ik heb wel eens trainingen gegeven aan mensen die in hun dagelijks werk met XML editors aan de slag moesten. Die vreesden dan het ergste, maar al heel snel snapten ze niet wat het probleem eigenlijk was. Een moderne XML editor – er zijn er inmiddels voldoende – lijkt erg veel op Microsoft Word of een andere

volwassen tekstverwerker, alleen biedt het wat extra mogelijkheden om metadata toe te voegen. Tags aangeven is dus absoluut niet moeilijk of ingewikkeld. Het is gewoon een kwestie van één stapje extra in de standaard werkmethode opnemen."
Nu toch de naam van Microsoft is gevallen, de rol van dit bedrijf noemt De Koning opmerkelijk. "Het bedrijf is ook nauw bij de SGML/XML User Group Holland betrokken en ik moet ze nageven: ze volgen de ontwikkelingen heel goed en ze zijn er nauw bij betrokken. We hebben een uitstekend contact met het bedrijf. Maar tegelijkertijd heb ik toch de indruk dat zij zichzelf zo langzamerhand een beetje buitenspel zetten door zo vast te houden aan een gesloten bestandsformaat. Office OpenXML (OOXML) is vooralsnog ook geen standaard geworden en dat zou toch een duidelijk signaal moeten zijn. Aan de andere kant snap ik hun dilemma heel goed. Er staan nu eenmaal grote commerciële belangen op het spel."

RSS en Atom

Erg interessant noemt De Koning ook de ontwikkelingen op het gebied van RSS en het Atom publishing protocol. "Dat is natuurlijk een mooi voorbeeld van de mate waarin XML inmiddels wordt gebruikt. In feite kunnen we stellen dat het gebruik van XML inmiddels zo wijdverbreid is, dat het als het ware aan het verdwijnen is. Onzichtbaar wordt, is misschien een betere uitdrukking.
We komen het zoveel tegen dat het heel gewoon is geworden. Veel mensen kennen RSS-feeds vooral van internet waar het een inmiddels veel gebruikte methode is om nieuwe informatie op websites of blogs automatisch naar je toe te laten sturen. In dat opzicht is het de opvolger van de push-technologie die we in de jaren negentig zagen opkomen, maar die het niet gered heeft.

Minder bekend is echter het gebruik van RSS binnen organisaties of – zoals de Nederlandse overheid bijvoorbeeld wil – als mechanisme om mensen en bedrijven de mogelijkheid te geven om officiële publicaties automatisch te ontvangen. Bedrijven als IBM en Microsoft werken daarnaast aan het gebruik van RSS als hulpmiddel om bedrijfsinformatie toegankelijk te maken, terwijl de overheid RSS als publicatiemechanisme ziet. Het past wat dat betreft ook in een groot architectuurproject binnen de overheid waarbij data eenmalig wordt vastgelegd bij één eigenaar: alle kadastrale gegevens bij het Kadaster, alle persoonsgebonden informatie bij de gemeenten, noem maar op. Is dat eenmaal gerealiseerd, dan kunnen dus ook nieuwe en op XML gebaseerde ontsluitingsmechanismen worden toegepast. RSS en Atom passen daar heel goed bij en beginnen – zoals het genoemde project rond de digitale staatscourant wel laat zien – steeds meer mainstream te worden. Net als XML dus."

Robbert Hoeffnagel is freelance journalist.

Update

Business Objects introduceert EPM XI

Business Objects brengt een volledig geïntegreerde suite van best-of-breed enterprise performance management applicaties EPM XI op de markt.
De suite biedt een nieuwe range van performance management oplossingen, ontworpen om grote business uitdagingen voor de CFO en operationele beslissers aan te pakken.
Business Objects bewerkstelligde binnen 5 maanden de volledige integratie van de Cartesis- en ALG-producten waardoor een combinatie ontstaat van tools voor financiën, performance management, risk governance, compliance en winstgevendheid samen met het bekende BI-platform van Business Objects. BusinessObjects EPM XI is naar eigen zeggen de eerste en enige in zijn soort: de EPM-suite incorporeert financieel intelligente mogelijkheden voor EPM, BI, data-integratie en datakwaliteit.
EPM XI gebruikt één geïntegreerd datamodel, IDM, dat de hoeksteen van de oplossing vormt en data-accuraatheid en -consistentie door alle EPM-processen waarborgt.

Business Objects heeft definitieve overeenstemming bereikt over de overname van het Duitse bedrijf FUZZY! Informatik. Het bedrijf is niet aan een beurs genoteerd; welk bedrag er met de overname gemoeid is werd niet bekend gemaakt. FUZZY! Informatik, gevestigd nabij Stuttgart, levert al meer dan 13 jaar datakwaliteitsoplossingen en hoort tot de grootste aanbieders in Europa. Het bedrijf is een spin-off van DaimlerBenz en heeft klanten zoals Daimler Chrysler, BMW, Cortal Consors, Deutsche Post, Mazda en O2. Het bedrijf heeft een sterke en bewezen oplossing in het behandelen en opschonen van Europese adresgegevens, en het creëren van directory content.

Met deze acquisitie verbreedt Business Objects zijn aanbod op het gebied van EIM (enterprise information management) met datakwaliteitoplossingen voor het cleansen van adressen in alle belangrijke Europese landen, integratiemogelijkheden met SAP en Siebel, on-demand adres-cleansing services, en verticale toepassingen voor de financiële industrie en postale services.