Metadata- en masterdatabeheer vormt de sleutel tot BI/PM succes

Semantic MediaWiki (2)

Henk Scholten

In het eerste artikel in DB/M 1 zijn de uitgangspunten en de opzet van een flexibele en complete architectuur voor Business Intelligence/Performance Management gepresenteerd. In dit artikel wordt dieper ingegaan op de praktische uitvoering daarvan.

Ter herinnering verwijzen we naar het overzicht van de architectuur in het vorige artikel, de afbeelding op pag 25. van DB/M 1. Deze architectuur is gebaseerd op duidelijke uitgangspunten en de aanbevelingen zijn concreet en onderscheidend. De drie belangrijkste uitgangspunten van deze architectuur zijn:

- Het bieden van alle toepassingen die samen een complete en geïntegreerde BI/PM-omgeving vormen;
- De realiteit van de uitvoering van BI/PM-projecten; de kritische succesfactoren, de meest voorkomende problemen en de typische organisatiedynamiek met betrekking tot BI/PM-projecten;
- Enige niet-functionele eisen; navolgbaarheid en het voldoen aan regelgeving (auditability & compliance).

De drie belangrijkste kenmerken van deze architectuur zijn:

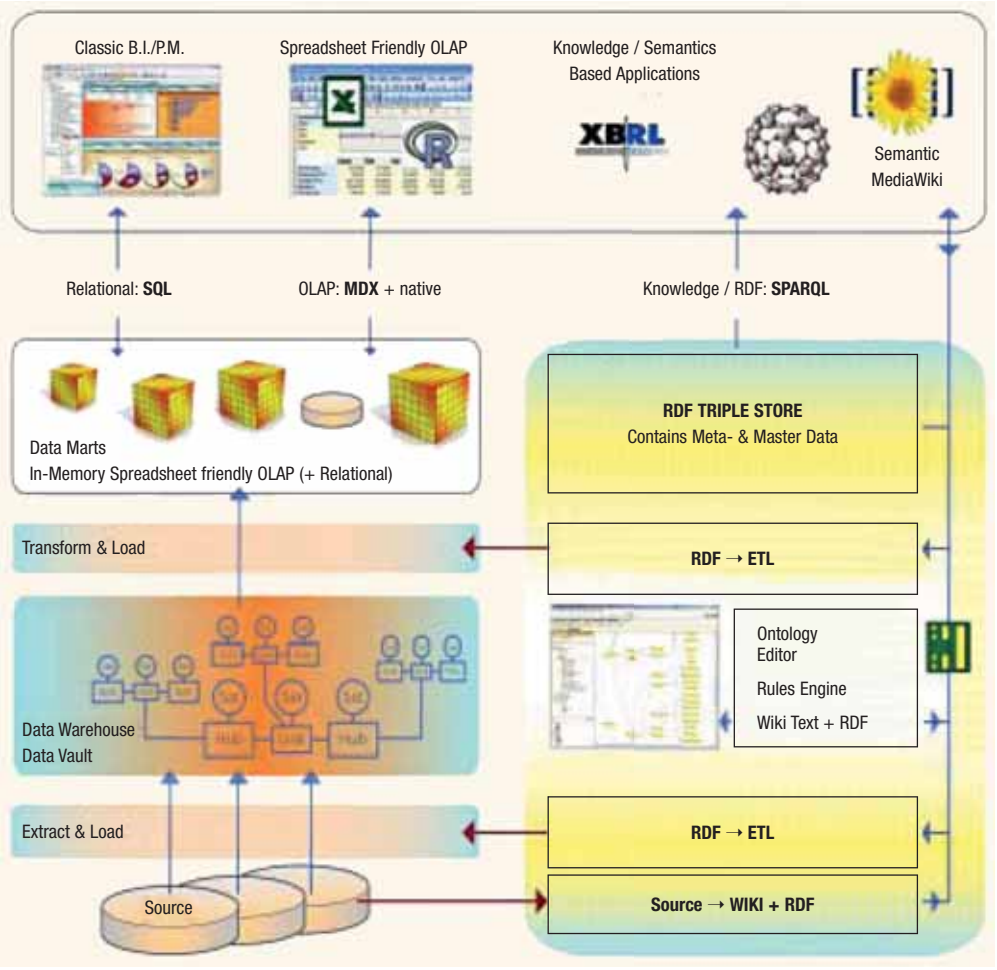
- Kennis en feiten worden duidelijk gescheiden opgeslagen van behoeften en waarheden (interpretaties) op basis van de Data Vault datawarehouse-strategie;
- Centrale positie voor in-memory OLAP-technologie met real-time read/write Excel-verbinding, met eenvoudig aan te sluiten rapportage en web-toegang;
- Metadata en masterdata management met een gestructureerde (semantische) Wiki als verbindende factor en voor het genereren van de ETL-code.

De voordelen zijn onder andere:

- Op basis van deze architectuur kan functionaliteit snel worden gerealiseerd. De toepassingen kunnen snel, flexibel en zoveel mogelijk door (super)gebruikers worden gerealiseerd. Een Agile/Scrum projectaanpak is optimaal uitvoerbaar met de voorgestelde componenten en opzet;
- Het kunnen ontdekken en volgen van de behoeften van de organisatie wordt optimaal ondersteund;
- De integratie van onder andere rapportage, analyse, forecasting, planning, budgetteren, kwaliteitscontrole, dashboarding, toezichtfuncties en beheer van niet-klassiek gestructureerde kennis voorkomt dubbel werk en inconsistentie;
- De inzet van bekende, breed toepasbare en toegankelijke software zoals Excel en (Semantic Media-) Wiki resulteert in optimale bruikbaarheid en controle door de gebruikers en lage kosten. Van de meest gebruikte BI/PM-tool, Excel, wordt met in-memory OLAP de meest bruikbare BI/PM-tool gemaakt;
- De aandacht is gericht op kennisbeheer, en niet op behoeftenanalyse. Vrijwel de gehele inspanning van de gebruikers en consultants tijdens de ontwikkeling blijft van direct nut. Op basis van de kennis in de organisatie worden het datawarehouse en de datamarts gebouwd en gevuld;
- Het op traditionele wijze realiseren van ETL-code kost veel consultancytijd, het genereren van die code brengt kostenbesparing en snellere realisatie van datastromen;
- De navolgbaarheid (auditability) en het voldoen aan regelgeving (compliance) zijn optimaal geborgd.

Het topje van de ijsberg

Bij BI/PM vormen de gebruikerstoepassingen het topje van de ijsberg. BI/PM drijft op data, dat wil zeggen: ETL-procesdefinities voor het bouwen en vullen van het datawarehouse en de datamarts. Verreweg het grootste deel van het werk en de onzekerheden zitten in het deel onder water, de datastroom en data-opslag. Als er goed wordt omgegaan met de behoeften van de gebruikers en van de organisatie, dan blijft data-ontsluiting over als het grootste risico bij BI/PM-projecten. De snel groeiende behoefte aan toegang tot real-time informatie, ongestructureerde informatie, externe data, enzovoort, maakt dat het realiseren van dit 'onzichtbare' deel van de BI/PM-omgeving voor IT een steeds grotere uitdaging vormt. In deze opzet hoeft IT zich bij de realisatie van het datawarehouse gelukkig maar marginaal met de wensen van gebruikers bezig te houden, alle relevante feiten worden 'RAW' in het datawarehouse opgeslagen. Er vindt dus bij het bouwen en vullen van het datawarehouse geen discussie plaatst over hoe de brondata bewerkt moeten worden. Dat helpt enorm bij de snelheid van implementatie. Immers, in een datawarehouse volgens de Data Vault-structuur staan "The Facts and nothing but the Facts". De behoeften (requirements) komen natuurlijk wel aan de orde, maar dan vooral bij de inrichting van de datamarts. En op datamart-niveau maakt de inzet van een in-memory OLAP server met een spreadsheet add-in als client een op 'requirements discovery' gerichte Agile/Scrum-aanpak heel goed mogelijk.



Afbeelding 1: Plaats binnen de architectuur van de applicatie die de Semantische Wiki en Triple Store inzet voor metadata en masterdata management plus ETL-generatie.

De sleutel tot succesvolle BI/PM

Dat de waarde van informatie staat of valt met kennis van wat de informatie betekent, lijkt een voor de hand liggende opmerking. Maar niet zelden is die kennis het grote struikelblok bij de realisatie van BI/PM-projecten en het gebruik van BI/PM. Metadata en masterdata zijn belangrijk in de gehele keten van de informatievoorziening. Vanaf het invoeren en vastleggen van de informatie, via de tussenstations, tot aan de interpretatie van de informatie door de gebruikers.

Om duidelijk te zijn over dit belangrijke onderwerp is ervoor gekozen om metadata en masterdata expliciet apart te benoemen, ook al zijn het verweven onderwerpen. Deze twee aspecten van kennisbeheer hebben elk hun eigen dynamiek en vragen om eigen beheer. Ook al lijkt het soms moeilijk te onderscheiden, het verschil is reëel, bijvoorbeeld in relatie tot consequenties van veranderingen. Het kan worden vergeleken met wat programmeurs Objecten en Instanties noemen, in de kennisbeheer (ontologie) wereld heet dit Tbox en Abox, het idee/concept en de realisatie.

Een beknopte omschrijving van metadata en masterdata is:

- Metadata betreft informatie over data (Object, Tbox, idee/concept). Deze beschrijven de data. Metadata definiëren bijvoorbeeld welke elementen samen een volledige en geldige beschrijving leveren van een object zoals een klant, product, leverancier, medewerker enzovoort. Metadata formuleren die beschrijving, maar beschrijven niet de concrete bestaande objecten of transacties. Veranderingen in de metadata hebben vrijwel altijd gevolgen voor de IT-bronsystemen; de databases en de schermen van softwaretoepassingen hebben hiermee een directe relatie;
- Masterdata zijn referentiedata (Instantie, Abox, realisatie). Ze beschrijven echt bestaande Sleutel Objecten en leveren de referentie voor de kennis over die concrete bestaande objecten. Ze kunnen bestaan uit een referentielijst van de klanten of producten met de belangrijke details of links naar details. Masterdata zijn direct verbonden met de in de organisatie gebruikte belangrijke informatiesleutels zoals product- en klantnummers.

Veel organisaties werken met meerdere informatiesystemen, het is dan extra belangrijk om te weten wat de 'gouden standaard' of het 'Nieuw Amsterdams Peil NAP' van de informatie is, en waar die staat. William (Bill) Inmon, Bonnie O'Neil en Lowell Fryman noemen dit Business Metadata in hun interessante boek met die titel [2]. Veranderingen in de masterdata liggen dicht bij de organisatie en het is wenselijk dat die snel kunnen worden doorgevoerd door personen die werken binnen de primaire processen

van de organisatie. Bijvoorbeeld: de clustering en eventueel hiërarchische indeling van klanten en producten moet door de organisatie kunnen worden veranderd, zonder dat dit specifieke IT-inspanning vraagt. Veranderingen in de masterdata kunnen gevolgen hebben voor 'IT dichtbij de gebruiker' zoals de opbouw van dimensies in OLAP-kubussen, maar niet voor de structuur van IT-bronsystemen en niet voor een Data Vault. Dus het systeem zelf en/of (super-)gebruikers kunnen de gevolgen van die veranderingen verwerken.

Centraal zenuwstelsel

Het opbouwen, onderhouden en delen van gemeenschappelijke kennis is onmisbaar voor effectief samenwerken. Als dat niet formeel en expliciet wordt gedaan, dan wordt het wel informeel en impliciet gedaan. Kennisbeheer vormt het centraal zenuwstelsel van de organisatie. Het staat vast dat de kwaliteit van de informatievoorziening direct afhankelijk is van het impliciete en expliciete kennisbeheer in de organisatie. Metadata en masterdata vormen daarvan de kern, maar het documenteren van procedures en het toegankelijk maken van ongestructureerde informatie horen ook bij kennisbeheer. Expliciet metadata- en masterdatabeheer is een zaak van de gehele organisatie, met een belang dat BI/PM, datakwaliteit en IT-implementaties ver overstijgt. Er komen veel aspecten kijken bij de inspanning tot het verkrijgen van een gemeenschappelijk inzicht en gedeelde informatievoorziening, om er enkele te noemen:

- IT. Alle IT-implementaties hebben direct te maken met de metadata binnen de organisatie, en ook in zekere mate met de masterdata;
- Bedrijfsmatig. De mensen in de organisatie moeten exact weten wat ze moeten en kunnen invoeren en hoe data moeten worden geïnterpreteerd. Referentielijsten van bijvoorbeeld klanten, leveranciers en producten zijn van groot belang voor het gecoördineerd en effectief werken;
- Politiek. Interpretaties en de aard van de exacte vastlegging van informatie dienen niet zelden een beperkt doel en komen niet noodzakelijkerwijze de gehele organisatie ten goede. Daarbij is unieke kennis niet zelden belangrijk voor posities. Dit maakt het vanuit de organisatie gezien juist extra belangrijk om systematisch met kennisbeheer om te gaan;
- De omgeving. Data worden uitgewisseld met vele belanghebbenden, onder andere de overheid, toezichtorganen, klanten, leveranciers en dienstverleners. Kennisbeheer is essentieel voor een goede uitwisseling van informatie, een op semantiek gebaseerd platform vormt een natuurlijke basis voor gestructureerde gegevensuitwisseling.

Kenmerkende dynamiek

In de wat grotere organisaties bestaan over het algemeen meerdere BI/PM-initiatieven gelijktijdig, de coördinatie daarvan is een bron van zorg. De term 'stove pipes' wordt graag gebruikt om aan te geven dat alle initiatieven gecentraliseerd zouden moeten worden. Maar de behoeften van diverse gebruikersgroepen veranderen voortdurend, en dat niet gelijktijdig en in dezelfde richting. Het gevolg is dat het BI/PM-landschap ondanks de mooie theorie in de praktijk vrijwel altijd een heterogene omgeving is. Die bestaat uit de resultaten van eerdere BI/PM-initiatieven en uit geïsoleerde oplossingen die specifiek bij bepaalde behoeftes passen. Planning is niet zelden volkomen gescheiden van rapportage, terwijl dezelfde data gebruikt worden. Gemeenschappelijk metadata- en masterdatabeheer plus een centraal datawarehouse waarin de feiten voor iedereen bruikbaar zijn opgeslagen kunnen een efficiënte bindende factor voor alle BI/PM-initiatieven vormen.
BI/PM is in eerste instantie geen *system of record* en wordt vaak ook niet zo opgezet. Maar planning, budgetteren en zelfs uitgebreide rapportage en analyses zijn producten waarvoor invoer plaats vindt die bewaard moet worden. Dat aspect moet niet uit het oog worden verloren. Voor een goed beheer is het ook belangrijk dat alle betrokken partijen zien dat BI/PM geen ERP, CRM of HRM is. Het draait bij BI/PM om verandering en wendbaarheid (agility), "The essence of BI/PM is Change not Run".
Bij het beheer van metadata en masterdata speelt verandering dus ook een belangrijke rol. Het in kaart brengen van metadata en masterdata is geen project dat ooit klaar is, maar een proces dat ingericht wordt.

Budget en Tijd

Metadata- en masterdatabeheer heeft pas echt nut als de gehele organisatie er bij is betrokken. Het functioneert pas goed als iedereen in de organisatie die informatie gebruikt en er aan bijdraagt. Maar grote *big bang* projecten hebben juist een geringe kans van slagen. Het beste kan daarom concreet met MDM worden begonnen ten behoeve van een specifiek BI/PM-project, dan staat de omgeving en bevat het alvast enige informatie.
Vervolgens moeten andere groepen gebruikers zonder barrières kunnen aanhaken. De MDM-applicatie mag daarom niet veel kosten. Gebruiksbarrières zoals beperkte licentieaantallen en nieuw aan te leren applicaties met hun gebruiksmetaforen moeten zo laag mogelijk zijn. Metadata- en masterdatabeheer mag niet als last worden ervaren, het moet direct nut opleveren voor wie gaat bijdragen. Spontane bijdrages en organische groei zijn dus veel acceptabeler dan een *big bang* aanpak.
Goed kennisbeheer kan organisaties veel voordeel opleveren, toch is het moeilijk om een budget beschikbaar te stellen. Het is ook ongrijpbaar wie welk voordeel heeft, hoe de kosten toegerekend moeten worden. Pas als de kosten laag genoeg zijn is het een gemakkelijk te aanvaarden stuk infrastructuur en hoeft er niet veel te worden overlegd over een verdeelsleutel. Een afdeling die het initiatief wil nemen kan de kosten van het inrichten van de hier voorgestelde MDM-omgeving al snel zelf dragen op één project en het dan aan iedereen ter beschikking stellen.
BI/PM-projecten zijn de aangewezen motoren van MDM-initiatieven. Juist doordat de kosten van BI/PM-projecten voor een belangrijk deel door de beschikbaarheid van de kennis over de data worden bepaald, kan in BI/PM veel worden gewonnen

met goed kennisbeheer. Het grote en groeiende nut voor de *gehele* organisatie komt daar als extra bij. Met het inrichten van een goede omgeving voor kennisbeheer, kan naast het nut voor de BI/PM-doelstellingen, veel extra waarde voor de organisatie worden gecreëerd.
Dat belangrijke kennis in de hoofden zit van mensen die beperkt en moeilijk beschikbaar zijn, is ook een belangrijk aspect. Het willen vastleggen van die kennis wordt al snel als belasting ervaren. De kennis moet dus op eigen gelegenheid en tijdstip gemakkelijk kunnen worden toegevoegd. Ook ziet men graag iets voor de inspanning terug, alleen vragen aan mensen om hun kennis af te geven werkt niet. Het zoeken is dus naar een omgeving waar vanuit de gehele organisatie voortdurend kennis aan kan worden toegevoegd en direct ook weer kennis uit kan worden gehaald: een omgeving die het opbouwen en delen van kennis bevordert. Daarbij mag het niet als 'project' door het leven gaan en mag het weinig tot niets kosten.

De Encyclopedie van de Organisatie

Er bestaat dus behoefte aan een 'Encyclopedie van de Organisatie' die door alle mensen in de organisatie kan worden onderhouden en geraadpleegd. Zoiets als Wikipedia, de bekende encyclopedie op het internet die op 15 januari 2001 is begonnen en precies acht jaar later 12,5 miljoen artikelen in 264 talen bevatte. Wiki Wiki is Hawaiaans voor snel, beweeglijk, en dat is precies waar het in BI/PM om draait. Zoals in afbeelding 1 is te zien vormt metadata en masterdata management een belangrijk en hét integrerende deel van deze architectuur. Temeer omdat de ETL-processen waarmee de Data Vault en datamarts worden gebouwd en gevuld, direct vanuit de op deze wijze vastgelegde kennis kunnen worden gegenereerd. De inzet van een gestructureerde Wiki maakt dat de IT-systemen ook met de informatie in de Wiki kunnen werken. Zo kunnen alle BI/PM-toepassingen gebruikmaken van de metadata en masterdata die in de database van de Wiki zijn opgeslagen.

AAA

In het Engels heet het dat de volgende drie A's kenmerkend zijn voor een Wiki: *Anybody can say Anything about Anything*. Dit heeft zich bewezen als basis voor een snelle groei met inhoud van hoge kwaliteit. De volledige openheid voor toevoeging heeft het Wiki-idee succesvol gemaakt. Discussiepagina's, een goed toegankelijke historie van de veranderingen en versiebeheer maken het ook gemakkelijk om de Wiki open te stellen.
Voor zover er kritiek op de inhoud van Wikipedia bestaat, is dit een gevolg van het volledig open en anonieme karakter. In Wiki's voor organisaties is dat anders, daar wordt vrijwel altijd met toegangscontrole gewerkt en kan wel iedereen zeggen wat hij wil, maar niet anoniem. Het maakt natuurlijk veel uit of de mensen die iets aan de Wiki toevoegen of veranderen, dat op anonieme basis kunnen doen, of dat de naam bekend is. In beide gevallen kan iedereen alles over alles zeggen, maar de spelregels en de betekenis van het woord 'alles' zijn duidelijk anders.

Intellipedia

Een goed en boeiend voorbeeld van de inzet van Wiki software voor operationeel kennisbeheer is het gebruik door de Amerikaanse inlichtingengemeenschap. Naar aanleiding van het falen van de CIA in relatie tot de gebeurtenissen van 9/11, werden haar werknemers aangemoedigd om hun ideeën voor de toekomst op papier te zetten. De inzending van Calvin Andrus met de titel "Toward a Complex Adaptive Intelligence Community" is zeker deels de inspiratie geweest voor de opzet van Intellipedia, de Intelligence Wikipedia. Het essay van Andrus is beschikbaar op het internet [1] en aanbevolen om te lezen. In 2005 werd een proefproject gestart en in april 2006 werd Intellipedia formeel aangekondigd. Intellipedia valt onder het 'Office of the Director of National Intelligence (ODNI), Intelligence Community Enterprise Services (ICES)' in Fort Maede, Maryland, USA. In april 2009 bevatte Intellipedia 900.000 pagina's die door 100.000 gebruikers waren gemaakt, met gemiddeld 5000 bewerkingen per dag [3].

In de uitspraken over de Intellipedia komt het belangrijke aspect van het verschil tussen anoniem en aangemeld gebruik duidelijk naar voren. Eén van de betrokkenen, Thomas Finger, maakt een vergelijking met eBay en zegt: "Intellipedia. It's been written up. It's the Wikipedia on a classified network, with one important difference: it's not anonymous. We want people to establish a reputation. If you're really good, we want people to know you're good. If you're making contributions, we want that known. If you're an idiot, we want that known too" [4]. Via diverse wegen is duidelijk gemaakt dat het een succes is, zie het Engelse Wikipedia artikel over Intellipedia voor referenties. Over Intellipedia stond in april 2010 een artikel in onder andere de Arabische, Siciliaanse en Russische Wikipedia, maar nog steeds niet in de Nederlandse. Intellipedia maakt, net als de concrete invulling van deze architectuur, gebruik van de open source MediaWiki software. Deze wordt door de groep achter Wikipedia gepubliceerd. De servers en zoekservices worden geleverd door Google.

Gestructureerde (Semantische) Wiki

De inzet van een gestructureerde (semantische, op betekenisleer gebaseerde) Wiki is een kenmerkend deel van de hier beschreven architectuur. Het woord semantiek staat voor betekenisleer, in een semantische Wiki heeft de computer toegang tot de betekenis van wat er in de Wiki-pagina's staat. De betekenis van wat op de Wiki-pagina's staat wordt voor de computer toegankelijk gemaakt door middel van het toevoegen van speciale labels aan de tekst; 'tags'. Een vaak genoemd resultaat van de inzet van semantiek is dat de kwaliteit van zoekresultaten sterk wordt verbeterd. Ook kan het gebruik van de moeilijk onderhoudbare Wiki-categorieën enorm worden beperkt, omdat een categorie wordt vervangen door een pagina met daarop een semantische query. Maar de echte betekenis zit dieper, want de Wiki wordt op deze manier opeens een natuurlijke user interface voor een

RDF Triple Store. Een Semantische Wiki kan gezien worden als een *belichaming van RDF*.
De semantische uitbreiding op de MediaWiki software, maakt deze uitermate geschikt voor kennisbeheer in een professionele omgeving De inzet van een semantische Wiki maakt deze architectuur onbedoeld geheel 'buzzword compliant', er mag een 'Web 3.0' sticker op de doos, want 'collaborative' wordt hier met 'semantic' gecombineerd.

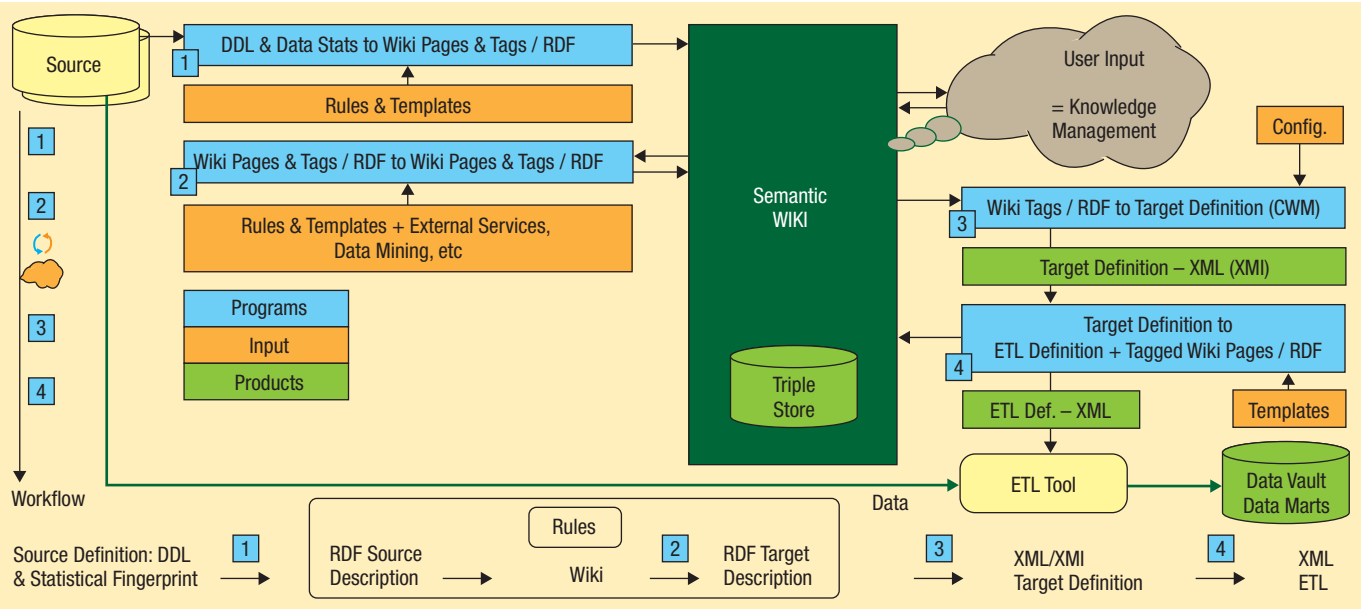
Resource Description Framework

Resource Description Framework (RDF) is een World Wide Web consortium (W3C) standaard, waarover W3C zegt: "RDF is a framework for supporting resource description, or meta-data (data about data)". Shelly Powers schrijft in zijn bij O'Reilly verschenen boek met de titel 'Practical RDF' het volgende: "RDF is used to capture specific statements about a resource, statements that help form a more complete picture of the resource." En dat is precies wat met metadata en masterdata management beoogd wordt te doen.
RDF berust op een heel eenvoudig concept. In elke taal kunnen eenvoudige feiten worden uitgedrukt met zinnen die elk bestaan uit drie specifieke stukken informatie: het onderwerp van het feit (in de Engelse RDF-terminologie: Subject); de eigenschap van het onderwerp dat met deze informatie wordt gedefinieerd (in RDF-termen: Predicate); de bijbehorende waarde (in RDF-termen: Object).
Bijvoorbeeld: Amsterdam (*Subject*) is de hoofdstad van (*Predicate*) Nederland (*Object*). Dit komt overeen met wat over het algemeen wordt ervaren als een volledige bewering, een complete eenheid van informatie. Met RDF kan een geheel beoogd domein beschreven worden door middel van zulke uitspraken, die elk een *triple* worden genoemd. Een triple is dus opgebouwd uit een Subject, Predicate en Object. De beschrijving van de kennis bestaat uit een (grote) verzameling van die korte 'zinnen'. Met RDF wordt de kennis in een consistente manier zo vastgelegd, dat die voor mensen en machines leesbaar en begrijpelijk is.
Het volgende beschrijft een RDF triple:

- Een (RDF) triple is een 3-tuple, dat bestaat uit een Subject, Predicate en Object;
- Elke (RDF) triple is een compleet en uniek feit;
- Elke (RDF) triple kan verbonden worden met andere RDF triples, maar het blijft zijn eigen unieke betekenis behouden, onafhankelijk van de complexiteit van het model waarin het is opgenomen.

Wiki-structuur en RDF

Een Wiki bestaat volledig uit pagina's. Die pagina's kunnen worden gezien als een 'proxy', een volmacht of vertegenwoordiger, voor een 'resource'; een reëel bestaand object, begrip enzovoort. Op een Wiki-pagina kan alles staan wat relevant is als informatie over het onderwerp van die pagina. Een Wiki-pagina is dus een natuurlijk Subject in RDF-termen, dat is intuïtief direct begrijpelijk. Voor RDF geldt dat het Subject uniek moet worden geïdentificeerd, de URL van een Wiki-pagina voldoet ook aan die eis. Er moet nu nog worden vastgelegd hoe de kennis over het Subject op de pagina wordt gecodeerd.
In elke Wiki bevatten pagina's links naar andere pagina's, die worden vastgelegd met het volgende eenvoudige label (tag): [[link_to_page]]. Door het opnemen van [[ ]] in de paginatekst, weet de MediaWiki software dat het iets speciaals met die tekst moet doen voordat er een webpagina van wordt gemaakt.
Wiki-pagina's die een aspect delen kunnen lid zijn van een categorie, bijvoorbeeld de categorie van hoofdsteden.
Dit wordt vastgelegd met een label volgens het patroon: [[Category:Name]]. Tussen de dubbele haken scheidt een dubbele punt het keyword Category van de naam van de bedoelde



Afbeelding 2: De BI-Team toepassing van Semantic MediaWiki voor BI/PM bestaat uit vier modules.

category. Met het Category label wordt een link gedefinieerd naar de categoriepagina met de naam: Name. Een dergelijke categoriepagina bevat automatisch links naar alle pagina's waarop een categorie-tag met deze Name staat.
In de gestructureerde (semantic) MediaWiki wordt één nieuw label toegevoegd: [[Predicate::Object]], let op de dubbele dubbele punt! Dit kan ook worden gezien als [[Semantic_Statement::Value]]. Hiermee wordt naast het doel van de link, ook de aard van de link vastgelegd. Er komen twee soorten semantische links voor, de Typed Link en de Attribute Tag. Oorspronkelijk werden die elk met een ander label patroon aangeduid, maar nu met dezelfde vorm [[::]]. Bijvoorbeeld in een pagina over de stad Amsterdam kan informatie vastgelegd worden met labels zoals een Typed Link naar een andere pagina: [[CapitalOf::the Netherlands]] en een Attribute met het label: [[population::750.000]]. In het eerste geval wordt verwezen naar een pagina over Nederland en is de relatie met die pagina vastgelegd met de tag CapitalOf. In het tweede geval wordt verwezen naar een pagina die het begrip population beschrijft. Dat is een belangrijk verschil. Het zou geen nut hebben om voor elk mogelijk inwonersaantal een aparte pagina voor dat getal op te nemen.
Samenvattend: op een Wiki-pagina waarmee een specifieke Resource wordt beschreven (het RDF Subject), wordt een label geplaatst met het patroon: [[Predicate::Object]].

Vrijwel elke vorm van kennis kan met RDF statements vastgelegd worden in een semantische Wiki. De semantische labels kunnen met de standaard Wiki editor worden bewerkt, maar ook met hulp van een Semantische Annotatie Editor of geautomatiseerd worden geplaatst in de *wikitext*; het eigen compacte opmaakformaat van MediaWiki. De grote bruikbaarheid van Wiki's is natuurlijk veel meer mensen opgevallen. Zo is bijvoorbeeld het bedrijf Vulcan van Microsoft medeoprichter Paul Allen actief als sponsor van het Semantic MediaWiki-project. Het bedrijf Ontoprise dat banden heeft met de Universiteit van Karlsruhe wordt door Vulcan betaald om de bruikbaarheid van de Semantic MediaWiki software te optimaliseren. Het open source product SMW+ is daarvan het resultaat.

RDF Triple Store

In deze architectuur functioneert de Wiki als front-end voor een RDF Triple Store. Een Triple Store is een specifiek database management systeem dat is gemaakt voor de opslag en manipulatie van RDF data. Technisch kan het een interface naar een relationeel database management systeem zijn. Bijvoorbeeld de Jena Triple Store draait op MySQL of Oracle. Maar er bestaan ook technisch volledig voor RDF gemaakte oplossingen zoals OntoBroker, die een veel betere performance kent dan Jena. SPARQL query's en een rules engine zijn belangrijke functionaliteiten van RDF Triple Stores.
RDF kan worden gezien als een 'driehoeksmeting van kennis' en vormt het hoogste niveau van abstractie waarop informatie kan worden gedefinieerd en opgeslagen. Bij een RDF Triple Store is er geen sprake meer van een databasestructuur die moet worden aangepast, alle RDF-informatie kan zonder meer worden opgeslagen. Eigenlijk zou het hele datawarehouse het beste in een RDF Triple Store kunnen staan. Helaas is dat nu nog niet reëel. Maar het door de Europese Unie gesteunde 'Large Knowledge Collider' project wil daarin verandering brengen. De Knowledge Representation and Reasoning Group van de VU is daarbij onder leiding van professor Frank van Harmelen betrokken. Deze complete architectuur kan in theorie zonder Wiki functioneren, alle query's lopen toch al op de Triple Store. De user interactie kan ook via formele kennis editors zoals Protégé of NeonToolkit verlopen, in aanvulling op een Wiki als user interface. Gegeven het abstractieniveau, kan een RDF Triple Store goed als centrale repository van metadata en masterdata dienen. Informatie in andere systemen voor metadatabeheer, zoals bijvoorbeeld Oracle Warehouse Builder, kan altijd in RDF-formaat worden uitgedrukt en/of vanaf RDF-notatie worden geladen.

SPARQL

Zoals SQL en MDX toegang geven tot relationele en multidimensionele databases, is SPARQL een door de W3C gestandaardiseerde querytaal die toegang geeft tot een Triple Store. SPARQL definieert ook een XML antwoordformaat en een webservice, elke SPARQL provider is een Service Oriented Application. SPARQL zoekt altijd naar patronen in de gehele database, bij het schrijven van de query's hoeft dus geen rekening te worden gehouden met een situatiespecifieke databasestructuur. De hier gepresenteerde toepassing gebruikt SPARQL op essentiële plaatsen. Over SPARQL is door Philip McCarthy een aanbevolen artikel verschenen met de titel: Search RDF data with SPARQL [6].

Kennisbeheer en ontologieën

In het eerdere voorbeeld werden de labels 'CapitalOf' en 'population' gebruikt, maar dat is dus taalgebonden, en wat is precies de uitleg? Effectieve samenwerking en communicatie vereist een gemeenschappelijk woordgebruik. Een ontologie bevat de beschrijving van de terminologie, concepten en relaties voor een bepaald kennisgebied. Een ontologie kan gezien worden als een normatieve en volledige beschrijving van de betekenis van de termen in een specifiek kennisdomein, daarmee is het de sleutel tot communicatie. Een expliciete beschrijving van de begrippen kan heel nuttig zijn voor mensen die samenwerken. Een ontologie bevat dus niet alleen feiten, maar ook regels die in formele logica zijn beschreven. Een ontologie bevat dus meer dan in RDF kan worden uitgedrukt, voor het beschrijven van een ontologie is bijvoorbeeld de W3C standaard OWL (Web Ontology Language) geschikt. Triple Stores kunnen aan de hand van een OWL-definitie ook afgeleide kennis als antwoord geven op een vraag. Bijvoorbeeld als vast staat dat X een *ChildOf* Z is en Y een *ChildOf* Z is, dan kan worden geconcludeerd dat X en Y verwant zijn. Als daarbij het geslacht van X en Y bekend is, kan worden vastgesteld of X en Y broer en/of zus zijn.

Het voorbeeld van Wikipedia laat zien hoe het werkt als ogenschijnlijk dezelfde begrippen in andere culturen worden gebruikt. Bij Wikipedia worden artikelen niet rechtstreeks vertaald, maar steeds weer wordt het begrip beschreven vanuit de beleving in die cultuur en context. Dat is een duidelijk model om met begrippen in een internationaal werkende organisatie om te gaan. Een Wiki biedt daarmee een bekende metafoor voor gebruik en kan goed dienen als referentie in een globale organisatie. Er bestaat een aantal verzamelingen van labels voor bepaalde informatiegebieden, die breed gebruikt worden, zoals *Dublin Core* en *FOAF* (friend of a friend). Wie op de juiste manier gebruik maakt van deze labels, maakt zijn informatie voor de hele wereld toegankelijk.
Het vanaf niets opstellen van een ontologie is geen eenvoudige zaak. Gelukkig zijn organisaties geen (volledig) open wereld en is het dus mogelijk om binnen een eindige hoeveelheid tijd een voldoende omvattende ontologie op te stellen. Vooral belangrijk is dat het kennisbeheer, het metadata en masterdata management, kan groeien en al functioneert terwijl er nog geen formele ontologie beschikbaar is. Gebruikers en computersystemen moeten informatie aan de Wiki kunnen toevoegen, die later automatisch of door gebruikers/redacteuren wordt voorzien van labels. Als er een ontologie is opgesteld, dan kan dat wel helpen bij het invoeren van nieuwe informatie omdat die dan meteen van de juiste labels kan worden voorzien. Het feit dat de ontologie voortdurend zal veranderen, hangt samen met de redenen waarom BI/PM-implementaties ook steeds veranderen. Het is belangrijk dat de ontologie leeft en steeds de organisatie volgt. Op basis van een ontologie kunnen Semantic Forms worden gemaakt, dat zijn Wiki-pagina's voor de structurele invoer van data.
Kennisbeheer en het genereren van de ETL-procesdefinities op basis van die kennis zijn kernelementen van deze architectuur. Dat maakt het mede mogelijk om wendbaar, 'agile', te werken en de behoeften op voorhand niet zo belangrijk te vinden. Belangrijk is ook dat de inspanning die van de gebruikers komt, direct rendement levert voor diezelfde gebruikers en de organisatie. En dat dus los van de doelstellingen op het gebied van technisch metadata- en masterdatabeheer. De in de toepassing ingezette Semantic MediaWiki+ software kan ook dienen als een vorm van document management, (media)bestanden kunnen aan artikelen worden gekoppeld.

Concrete toepassing Semantic MediaWiki

De toepassing genereert Wiki-pagina's voor elk object van de brondatabases (tabellen en velden) en zet de beschikbare informatie over die objecten als labels op die pagina's. Dit wordt dan bij het opslaan van een Wiki-pagina automatisch direct in de RDF Triple Store opgeslagen. Vervolgens kan de kennis uit de organisatie op eenvoudige wijze aan de Wiki-pagina's worden toegevoegd. Op grond van deze informatie wordt intelligent besloten hoe de Data Vault en datamarts eruit gaan zien. Dit is een iteratief proces, waarbij informatie door de gebruikers en de *rules engine* aan de pagina's wordt toegevoegd. Totdat er voldoende informatie beschikbaar is op basis waarvan de *rules engine* het beoogde resultaat kan produceren. Vervolgens worden de definities gegenereerd waarmee een ETL-tool de Data Vault en datamarts kan bouwen en vanuit de bron laden. En er worden Wiki-pagina's gemaakt met informatie over de Data Vault en datamarts.
De toepassing bestaat uit vier modules, die na elkaar worden ingezet en elk het product van de vorige fase als input gebruiken, zie afbeelding 2. De producten van deze applicatie bestaan uit Wiki-pagina's (content) en configuratiebestanden voor een ETL-tool (code). De programma's die lopen zijn bewezen en breed bekend, de inzet van deze toepassing is daarmee zonder risico. De producten van de vier modules kunnen met een standaard editor worden bekeken, bewerkt en desnoods gemaakt, voordat ze als input voor de volgende fase dienen. De software is in Groovy/Java geschreven en maakt gebruikt van Groovy Templates. Het concept werkt omdat een Data Vault en op basis daarvan de datamarts goed met behulp van templates gegenereerd kunnen worden. Voor de Wiki wordt de Semantic MediaWiki software ingezet:

1. In de eerste fase worden Wiki-pagina's gemaakt vanuit de Data Definition Language van de brondata. Zo mogelijk wordt een statistische vingerafdruk van de data op de Wiki-pagina's opgenomen. De pagina's bevatten dus alle beschikbare informatie over de brondata, die semantisch zijn gelabeld (van tags voorzien);
2. De tweede fase bestaat uit een iteratief proces waarbij software en gebruiker samen komen tot voldoende informatie op de pagina's om de Data Vault of datamarts te kunnen definiëren. De computer helpt de gebruiker om zijn aandacht te richten op pagina's die gebruikersinvoer vragen;
3. In de derde fase worden de labels die de Data Vault of datamarts definiëren gelezen en wordt daar een XML bestand van gemaakt dat de definitie (van delen) van de Data Vault of datamarts bevat. Die definities zouden theoretisch met de hand kunnen worden gemaakt en alleen bestaan uit de vermelding van brontabel en kolom, dit zou dan vanuit de Wiki kunnen worden verrijkt met alle veldinformatie zoals datatype en veldlengte;
4. De vierde fase neemt een XML Data Vault of datamart-definitie als input en levert XML bestanden voor een ETL-tool als product, en maakt desgewenst Wiki-pagina's die de Data Vault of datamarts beschrijven.

De tweede fase is duidelijk het meest interessant. Hier wordt de informatie vanuit de database en vanuit de gebruiker verwerkt tot conclusies. Daarvoor moet de informatie die van de gebruikers komt van labels worden voorzien. Er bestaat een keuze aan mogelijkheden om dat te doen:

- Uiteraard kan een gebruiker de labels direct ingeven als hij de pagina maakt;
- Eventueel kan een Semantische Editor helpen. Als er een ontologie is gedefinieerd, kan de Semantische Editor zorgen voor consistentie bij het invoeren van de labels;

- Interessant is de mogelijkheid om de tags automatisch aan te brengen of veranderen. Dat kan met een eigen tagging service of bijvoorbeeld met de OpenCalais webservice van Thomson Reuters, waarmee teksten in het Engels, Frans of Spaans automatisch van bepaalde tags kunnen worden voorzien.

Technisch gezien voert module twee een SPARQL query uit op de Triple Store en haalt een aantal kandidaat Subjecten op met een aantal Property's. Dit is de input voor een rules engine. De output bestaat uit labels op Wiki-pagina's die worden toegevoegd of veranderd, en eventueel nieuwe Wiki-pagina's met labels. Als rules engine wordt DRools van het JBoss-project gebruikt. Deze rules engine is breed toepasbaar binnen organisaties, onder andere voor Complex Event Processing (CEP). De BI-Team toepassing werkt met alle rules engines die voldoen aan de Java Rule Engine API standaard JSR-94, zoals bijvoorbeeld JESS+fuzzyJESS.

De DRools regels kunnen op diverse manieren worden opgesteld en onderhouden. Ze kunnen in een Domain Specific Language worden gedefinieerd en als goed leesbare tekst op Wiki-pagina's staan, of worden beheerd met de Guvnor web-based tool. En er kan ook met in Excel opgestelde decision tables worden gewerkt. Voor het concept is de inzet van een rules engine belangrijk, en de rules zelf bevatten belangrijke informatie. Module twee is voor veel meer bruikbaar dan alleen het bouwen van Data Vaults en datamarts!

Eindeloos veel mogelijkheden

Denk eens aan een Wiki-pagina met de actuele informatie voor elke klant, leverancier, product, concurrent enzovoort. Visuele en statistische analyses die met R worden gemaakt, kunnen in die Wiki-pagina's worden opgenomen. Ook rapporten die met BIRT van Actuate zijn gemaakt kunnen in Wiki-pagina's worden opgenomen. Ook real-time informatie die van web services wordt betrokken kan worden opgenomen. Een denkbaar voorbeeld is een pagina per belangrijke klant of product (masterdata), die regelmatig wordt aangevuld en bij de tijd wordt gehouden met informatie vanuit de organisatie, een knipseldienst, kredietinformatie, de Kamer van Koophandel enzovoort.
Met module twee kan de actuele inhoud van de Wiki voortdurend of op gezette tijden worden geëvalueerd, en de resultaten daarvan kunnen op (dezelfde) Wiki-pagina's worden gepubliceerd. De webservice-capaciteiten van Semantic MediaWiki+ en communicatiekanalen zoals RSS, maken die informatie beschikbaar voor mens en machine, de medewerkers en de computersystemen.
Het is bijvoorbeeld ook denkbaar dat de Nederlandse politie alle processen verbaal in een Semantische Wiki opslaat en dat manueel en automatisch gaat taggen, daarbij kunnen rules lopen voor real-time evaluatie. Met het goed ontsluiten van de nu al beschikbare informatie kan ongetwijfeld meer aan het bestrijden van criminaliteit worden bijdragen, dan met het verzamelen van steeds meer informatie.

Tenslotte

In dit artikel en het vorige is een complete architectuur gepresenteerd die wel gebruik maakt van specifieke technieken, maar niet is gebonden aan een leverancier. De elementen zijn beproefd en de werking is goed bekend. Belangrijke delen van deze architectuur worden al jaren met groot succes ingezet bij organisaties variërend van de grootste ondernemingen tot aan informatie-intensieve eenmanszaken. De 'credibility' van de in-memory OLAP-techniek is recent toegenomen doordat IBM/Cognos nu *in-memory* OLAP met TM1/Xcelerator volledig heeft geadopteerd en er meerdere leveranciers zijn bijgekomen. Maar de eerste resultaten daarmee werden in Nederland nu al ongeveer 15 jaar geleden door het bedrijf Metaldata gerealiseerd bij organisaties zoals ABN-AMRO, Heineken, Nederlandse Spoorwegen enzovoort. Data Vault is in korte tijd een in Nederland breed geaccepteerde datawarehousing-opzet geworden. En de Semantische MediaWiki werd door Gartner in een recent webinar over MDM genoemd als "very exciting" en veelbelovend voor metadata- en masterdatabeheer, met daarbij de leverancier Ontoprise op de kaart. De MediaWiki software kent met Wikipedia miljoenen gebruikers. Statistische analyse en complexe visualisatie zijn met RExcel en Rapid Miner, of SPSS en andere producten binnen handbereik van iedereen. De hier voorgestelde aanpak van BI/PM is geen avontuur, maar beproefd en nu volledig uitvoerbaar.

Referenties

Voor iedereen die in (Business) Intelligence is geïnteresseerd, eigenlijk voor alle managers, is het artikel van D. Calvin Andrus van belang.

1. *D. Calvin Andrus, Toward a Complex Adaptive Intelligence Community, The Wiki and the Blog, (15 april 2007), CIA, Center for the Study of Intelligence – CSI, https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol49no3/html_files/Wik_and_%20Blog_7.htm*
2. *William (Bill) Inmon, Bonnie O'Neil en Lowell Fryman, Business Metadata – Capturing Enterprise Knowledge, Morgan Kaufmann 2008, ISBN 978-0-12-373726-7*
3. *Massimo Calabresi, (8 april 2009),Time, Wikipedia for Spies: The CIA Discovers Web 2.0, www.time.com/time/nation/article/0,8599,1890084,00.html*
4. *Clive, Thompson (3 december 2006) , the New York Times, Open-Source Spying, http://www.nytimes.com/2006/12/03/magazine/03intelligence.html?pagewanted=print*
5. *Mark, Mazetti (12 april 2007), the New York Times. "Intelligence Chief Announces Renewed Plan for Overhaul", www.nytimes.com/2007/04/12/washington/12intel.htm*
6. *Philip McCarthy, Search RDF data with SPARQL, www.ibm.com/developerworks/xml/library/j-sparql/*
7. *Henk Scholten, BI-Team whitepaper, 2009, BI-Team Referentie Architectuur, Een compleet raamwerk voor Performance & Knowledge Management.*

Op de website van DB/M vindt u de volledige integrale tekst van de samengevoegde twee artikelen. Daarin zijn ook actieve links naar referentie websites te vinden.

Henk Scholten (hscholten@bi-team.com) is Managing Director van BI-Team BV.