Nieuw licht op een belangrijk probleem kan hopelijk inspiratie geven

Documentatie op basis van metadata (3)

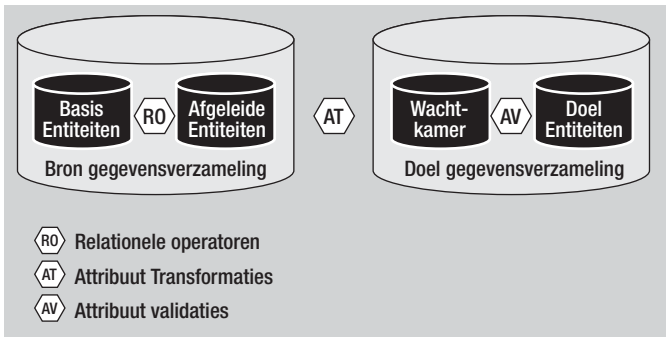
Burkhard Lau

In deel 2 is een beeld geschetst van documentatie, die in het ideale geval allemaal gestructureerd in een repository is opgeslagen. In dit laatste artikel in de serie gaan we na welke structuur het repository zou moeten hebben. Met deze kennis in het achterhoofd worden ook nog wat misverstanden ontzenuwd over metadata.

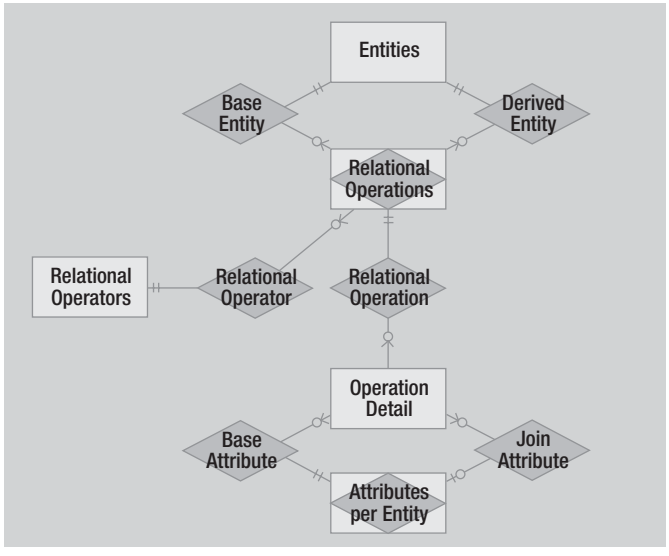
Een werkende DWH-omgeving levert als resultaat van informatiestromen de volgende gegevenssoorten op: bewerkte kopieën van bedrijfsgegevens; en procesgegevens. Terwijl het ontsluiten van bedrijfsgegevens de voornaamste taak van een DWH-project is, representeren zich de nodige ETL-processen, maar ook query's op de database en overige processen, om gegevens te kopiëren door hun procesgegevens. Door deze procesgegevens te koppelen aan de kopieën van bedrijfsgegevens kunnen vragen over de actualiteit van de gegevens of mutaties in het verleden worden beantwoord. Echter, ook het tunen van een database of de controle van de ETL-processen maakt de aanwezigheid van deze procesgegevens noodzakelijk.

De interpretatie van deze gegevens gebeurt in vier abstractieniveaus:

- conceptueel, procesonafhankelijk;
- functioneel, implementatie-onafhankelijk;
- technisch, implementatie-afhankelijk;
- operationeel, gebouwde gegevensstructuren en proces-software in de proceslaag.

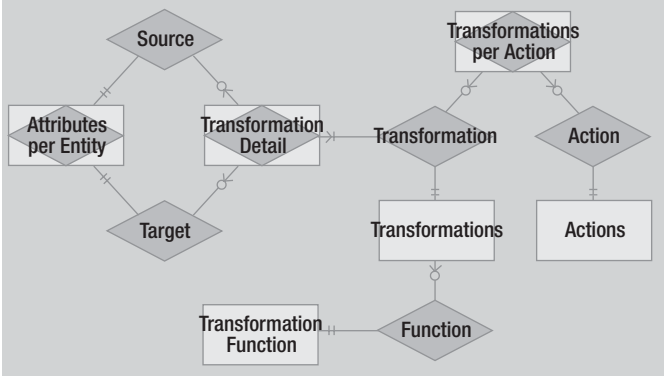


Afbeelding 1.



Afbeelding 2.

De eerste drie abstractieniveaus vormen samen de nodige documentatie, waarvoor we de benodigde opslagstructuren moeten gaan bepalen. Omdat we de documentatie als gegevens in



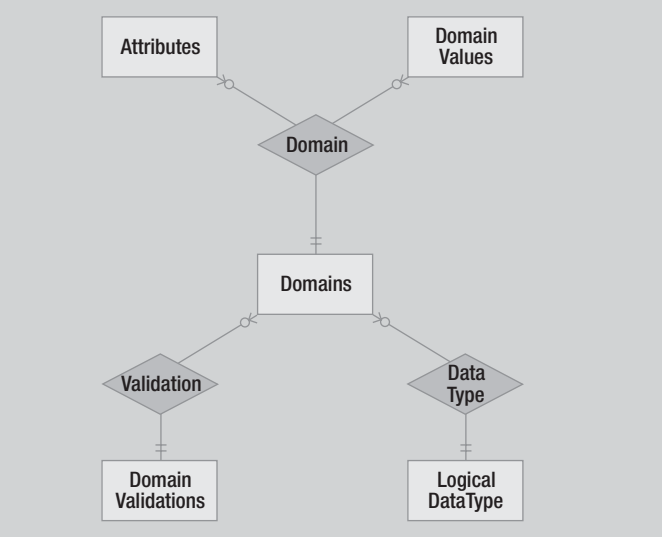
Afbeelding 3.

een database, dus als data, gaan opslaan, gaan we dit vanaf nu metadata noemen.

Structuur van de metadata

Een informatiestroom is generiek te beschrijven als een volgorde van steeds weer dezelfde structuur, zie afbeelding 1. Uitgaande van basisentiteiten (tabellen en views) worden via relationele operatoren (selectie, projectie, join, union) entiteiten afgeleid. Zie afbeelding 2 voor een ERD, dat dit proces in opslagstructuren vertaalt.
De attributen van deze afgeleide entiteiten worden vervolgens via transformatiefuncties (aggregaties, transformaties) in andere attributen omgezet (afbeelding 3). Als er sprake is van een validatie loopt dat via de 'wachtkamer', anders worden de doelentiteiten direct gevuld. Tijdens de validatie wordt op basis van het attribuutdomein de attribuutwaarde gevalideerd (afbeelding 4), en in geval van goedkeuring komt het attribuut terecht in de doelentiteiten. Wanneer echter de validatie niet slaagt, komt het betreffende record in een uitvaltabel terecht, die deel uitmaakt van de wachtkamerentiteiten. Gebruikelijk en voldoende is het om een attribuut op zijn weg vanuit de bron naar diverse rapporten slechts één keer te valideren.
Dit basisproces kan zich tussen twee verscheidene gegevensverzamelingen afspelen, of binnen één gegevensverzameling (zoals het vormen van een aggregaattabel), of binnen één entiteit (als een attribuut afgeleid wordt van andere attributen).
Voor de structuur van de metadata zijn nu twee aspecten interessant:

- Het wat: welke informatie stroomt, belangrijk voor logische datamodellering en interessant voor impact-analyses (afbeelding 5). Afwijkend van de voor de hand liggende modellering is in onze praktijk gebleken dat er vaak attributen zijn, die bij meerdere entiteiten voorkomen. Dit kunnen standaard attributen zijn, zoals een wijzigingsdatum, of attributen, die zowel in basisentiteiten als afgeleide entiteiten voorkomen;



Afbeelding 4.

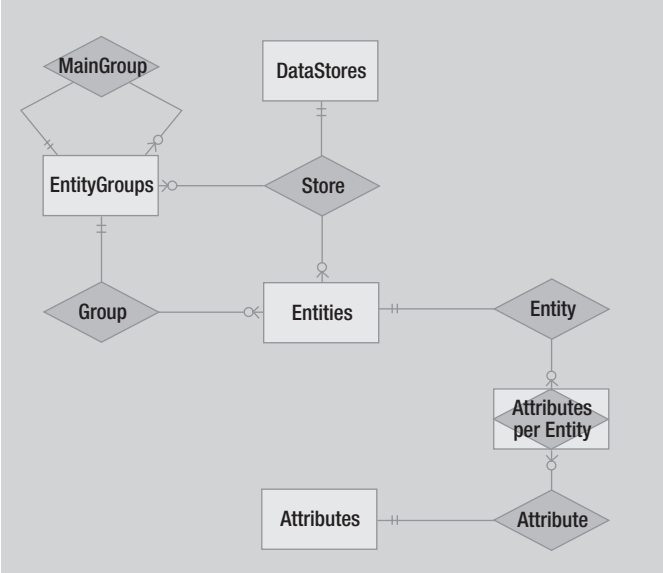
- Het hoe: hoe stroomt de informatie. Dit is belangrijk voor FO's, waarin de processen beschreven worden, zie afbeelding 6. Een belangrijk detail in het ERD is dat processtappen elkaar kunnen aanroepen, of via een externe scheduler in een procesvolgorde gezet zijn.

De praktijk

Om te controleren of de hier beschreven structuur wel in de praktijk voldoende is, zijn de sjablonen van een aantal van klanten tegen de structuur gehouden. In de onderaan de pagina staande tabel wordt een aantal vaak voorkomende sjabloononderwerpen vergeleken met de onderwerpen, die door de metadata afgedekt worden.
In deze praktijk-case kunnen we alle rubrieken van het sjabloon vanuit onze metadata vullen, waarbij de volgende opmerkingen van toepassing zijn:

- Het sjabloon voorziet niet in UNION en PROJECTIE

Sjabloon-item	Metadata-aspect	Opmerking
Ophalen	Basisentiteiten	De lijst van alle gebruikte bronentiteiten.
Verrijken	Relationele operator: Join	De gebruikte join om de bronbestanden te combineren.
Selecteren	Relationele operator: Selectie	De restrictie(s), die toegepast dienen te worden.
Koppelen	Attribuuttransformatie: join	Het specifieke geval van een attribuuttransformatie, waarbij niet de waarde getransformeerd wordt, maar de attribuutcombinatie gebruikt wordt om bron en doel te koppelen.
Valideren	Attribuut Validatie	
Foutafhandeling	Attribuut Validatie	
Vermeerderen	Attribuut Transformatie: 1-op-n relaties	
Manipulatie: toevoegen	Attribuut Transformatie per Actie	De attribuuttransformaties, die bij het toevoegen van record uitgevoerd moeten worden.
Manipulatie: bijwerken	Attribuut Transformatie per Actie	De attribuuttransformaties, die bij het bijwerken van record uitgevoerd moeten worden.
Manipulatie: verwijderen	Attribuut Transformatie per Actie	De attribuuttransformaties, die bij het verwijderen van record uitgevoerd moeten worden.



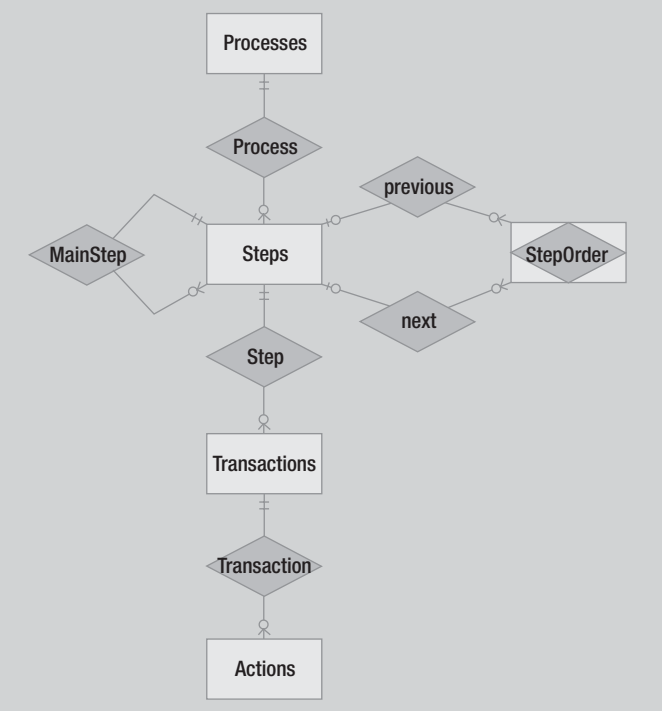
Afbeelding 5.

operatoren, dus óf het sjabloon zou moeten worden aangepast, óf deze twee operatoren mogen niet gebruikt worden bij het specificeren (als richtlijn te stellen voor deze klant);
- Het sjabloon houdt geen rekening met attribuutdomeinen, waardoor op de conventionele manier voor elk attribuut de validatie opnieuw beschreven moet worden;
- Hetzelfde geldt voor de foutafhandeling: is eveneens op attribuutdomeinniveau beschreven als onderdeel van validatie.

Metadata Misverstanden

Het begrip metadata wordt vaak lukraak toegepast. In het opleidingsmateriaal van een grote BI-dienstverlener zijn bijvoorbeeld de volgende vragen opgesomd die door metadata beantwoord zouden worden:

1. Hoe zit een transformatie in elkaar;
2. Is het bronsysteem gewijzigd;
3. Is het laadproces geslaagd;
4. Is de responstijd acceptabel;
5. Is de beschikbaarheid in orde;
6. Hoe recent zijn de data;
7. Welke productgroepen zitten er in;
8. Hoe is de omzet berekend?



Afbeelding 6.

Laten we kijken welke gegevens voor het beantwoorden van bovenstaande vragen nodig blijken te zijn, en of we ook inderdaad over metadata kunnen spreken. Een verkenning op Internet levert al gauw honderden definities op, waarvan hier twee voorbeelden:

Broadly, data about data, or information about information.
In practice, metadata comprises a structured set of descriptive elements to describe an information resource or, more generally, any definable entity.
(www.ktweb.org/rgloss.cfm).
Information about data; more specifically, information about the meaning of other data.
(www.nima.mil/vpfproto/vpfgloss.htm).

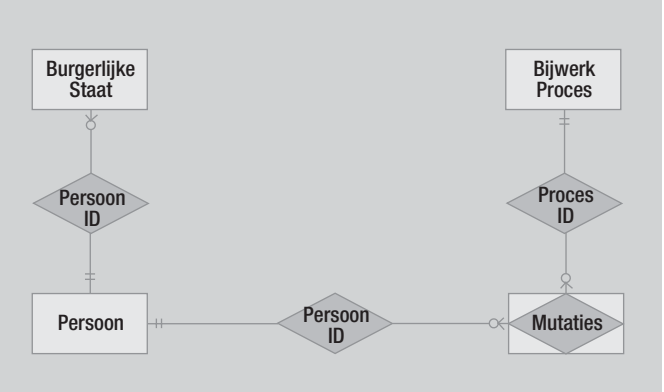
Letterlijk vertaalt zijn metadata gegevens over gegevens. Problematisch is echter om scherp te formuleren, wat het 'gegeven over een gegeven' precies betekent. Bekijk het onderstaande simpele datamodel.

Tabel	Veld	Toelichting
Persoon	ID	Identificatie v/d persoon
	Naam	De naam v/d persoon
	Geboortedatum	Geboortedatum v/d persoon
	Bijgewerkt	Het tijdstip dat voor het laatst gegevens van deze persoon toegevoegd of gewijzigd werden.
BurgStaat	Persoon	De identificatie van een persoon
	Van	Het begin van een periode
	Tot	Het eind van een periode
	Staat	De omschrijving van de burgerlijke staat v/d persoon binnen de periode.

De naam en het geboortedatum zeggen iets *over* de persoon, die door het record gemodelleerd wordt. Het veld Bijgewerkt zegt direct iets over de actualiteit van het record, en hiermee iets over het proces dat nodig is om dit record bij te werken. Als de uitvoering van dit proces door een log-regel in de database vertegenwoordigd was, zou ook een referentie naar deze log-regel voldoen voor dezelfde informatie.
De tabel BurgStaat tenslotte bevat de geschiedenis van de burgerlijke staten, dus informatie *over* de in tabel Persoon gemodelleerde personen. We hebben dus diverse informatie over iets, maar in geen enkel geval kunnen we van metadata spreken. Bij de persoonsgegevens (naam, geboortedatum, burgerlijke staat) hebben we te maken met gegevens, die een deel van de werkelijkheid weergeven, het mutatiegegeven daarentegen is een gegeven van het bijwerkproces, en zou ook zodanig gemodelleerd kunnen worden, zie afbeelding 7.
Zijn er dan helemaal geen metadata in ons voorbeeld? De kolom toelichting in de tabel bevat wel expliciete metadata, maar er zitten ook een heleboel impliciete metadata in het model. Niet gemerkt? U heeft toch vast en zeker bij het doornemen van de attributen niet expliciet genoemde eigenschappen aangevuld, zoals het datatype, de primaire sleutel en de referenties naar de persoonstabel? Hiermee hebben we een aardige voorzet voor een metadata-definitie te pakken: expliciete of impliciete beschrijvingen van gegevensstructuren of processen op een gestructureerde (repository) of ongestructureerde (documentatie) manier.

Geen Metadata

Duidelijk is geworden dat procesgegevens geen metadata zijn, maar gegevens die de verwerking van bedrijfsgegevens documenteren. Zij zijn geen abstractie van de bedrijfsgegevens, maar geven aanvullende informatie, die bij behoefte ook samen met de bedrijfsgegevens kan worden opgeslagen. Hiermee is bovengenoemde lijst op te delen in wel/geen metadata op de manier in onderstaande tabel:



Afbeelding 7.

Conclusies

In dit laatste artikel van de reeks over een metadata-gebaseerde documentatie/aanpak, hebben we ons over structuurdetails van een metadatabase gebogen. We hebben vervolgens gezien, dat procesgegevens geen metadata zijn, dat hun beschrijving echter als procesbeschrijvingen wel onderdeel uitmaakt van de metadata. Dit was de afsluiting van een reeks die misschien voor u een nieuw licht op een belangrijk probleem werpt, en hopelijk inspiratie geeft om ook in uw omgeving nog eens stil te staan bij de manier hoe met documenten wordt omgegaan, de specifieke problematieken expliciet en bespreekbaar te maken, en wellicht deze artikelen als aanleiding te nemen om structurele oplossingen te implementeren.

Burkhard Lau
Dr. Ir. Burkhard Lau (burkhard@keper.nl) is senior BI consultant bij BI Garant BV.

Wel Metadata	Geen Metadata	Toelichting
Hoe zit een transformatie in elkaar?		Dit betreft metadata in de zuiverste zin van dit artikel.
	Is het bronsysteem gewijzigd?	Als de vraag op de brongegevens betrekking heeft, is de vraag met de procesgegevens te beantwoorden. Als de vraag zich op de structuur van de bron richt, is naast de metadata ook een vergelijking met de bron nodig. Metadata in onze zin kan hier wel helpen, maar zijn hier voor niet primair bedoeld.
	Laadproces geslaagd?	Een typisch procesgegeven.
	Responstijd acceptabel?	Gebaseerd op procesgegevens en referenties kan een monitor deze vraag beantwoorden.
	Beschikbaarheid in orde?	Een typisch procesgegeven.
	Hoe recent zijn de data?	Een typisch procesgegeven.
	Welke productgroepen zitten er in?	Evenmin metadata, maar een gewoon bedrijfsgegeven, dat uit de bedrijfsdatabase op te vragen valt.
Hoe is omzet berekend?		Dit slaat weer wel op een beschrijving van een gegeven of een proces.