DATAMODELLERING TOEPASSEN DATA QUALITY

Inleiding

In dit whitepaper wordt een toepassingsgebied beschreven voor datamodellering. Een toepassing is een werkveld op het vlak van architectuur of modellering waarbij een aantal data modelleervormen met elkaar gecombineerd worden.

Deze specifieke modelleervormen zijn beschreven in een serie whitepapers. In de whitepapers over toepassingsgebieden gaan we in hoe de verschillende modelleervormen met elkaar gecombineerd worden ter ondersteuning van dit toepassingsgebied

Deze combinatie maakt het vervolgens mogelijk om op adequate wijze een model te communiceren voor dit toepassingsgebied. In een aantal gevallen wordt alleen documentatie geproduceerd, in andere situaties kunnen ook andere zaken geproduceerd worden zoals source code of templates etc.

Doel

Data Quality is een onderdeel van Data Management. Data Quality is een werkveld dat binnen elke organisatie waar data wordt verwerkt relevant is. Data met een lage kwaliteit veroorzaakt een aantal knelpunten in de bedrijfsvoering die naast financiële gevolgen ook gevolgen kan hebben voor het imago van de organisatie.

Wordt data geproduceerd, verwerkt, opgeslagen en getransporteerd dan is Data Quality relevant. Binnen deze activiteiten worden er bewerkingen gedaan op data waarbij de structuur van deze data relevant is, maar ook de kwaliteit van deze data binnen het gebruik. Zijn de gegevens accuraat, actueel, volledig enzovoorts. Wordt niet aan de kwaliteitseisen voldaan dan zal de organisatie moeten zoeken naar manieren in haar bedrijfsprocessen, zoals controle en correctiestappen, om de kwaliteit naar een voldoende hoog niveau te brengen.

Data modellering en Data Quality lijken in eerste instantie weinig met elkaar gemeen te hebben, echter niets is minder waar. Enerzijds is de waarde van de data gelegen in de structuur van de data en de mogelijkheid om de structuur naar behoefte te veranderen. De structuur van de data is daarmee een kwaliteitskenmerk geworden Anderzijds is data voor steeds meer organisaties een productiemiddel en heeft het daarmee waarde voor de organisatie en haar omgeving. Heeft data als productiemiddel een hoge waarde dan kan een organisatie zich daarmee onderscheiden ten opzichte van organisaties die werken met data van een lagere kwaliteit.

In dit whitepaper behandelen we de Data Quality vanuit het perspectief van het DaMa International, het consortium achter de DaMa Body of Knowledge, waarbij we regelmatig verwijzen naar achtergrond informatie. De hoofdpagina voor deze site is te vinden via http://dama.org

Context

Data Quality introduceren is wat reeds binnen alle organisaties plaatsvindt. Dat kan impliciet zijn. Vraag je bijvoorbeeld aan medewerkers wat is de kwaliteit van een bepaalde dataset dan kan men veelal perfect aangeven wat de kwaliteitsissues zijn.

Structureel processen uitwerken rond data kwaliteit is minder gemeengoed. Echter organisaties die hier tijd en effort in steken zullen na verloop van tijd besparingen bereiken op het vlak van tijd en geld omdat de bedrijfsprocessen optimaler zijn en het nemen van beslissingen gebeurd op basis van data met voldoende kwaliteit.

Data Quality is een onderdeel van Data Management. Data Management is een aantal aan elkaar gerelateerde bedrijfsprocessen met focus op de diverse aspecten van data. In het DaMa Body of Knowledge wordt het onderstaande raamwerk voor deze bedrijfsprocessen uitgewerkt

Bron: DMBoK

Dataverwerking wordt bij veel organisaties steeds complexer. Enerzijds doordat de structuur van de datasets steeds complexer wordt. Anderzijds doordat organisaties steeds meer data verzamelen. Als laatste zijn de nieuwe mogelijkheden rond data analytics en data science te noemen.

Heeft een organisatie hierbij de keuze uit datasets met verschillende kwaliteitsniveaus dan helpt dit om optimale configuraties te kiezen voor bovengenoemde toepassingen.

DOELEN VAN DATA QUALITY

Doelen van data quality zijn met name gericht op het definiëren van data kwaliteit en het nemen van maatregelen die bijdragen aan data met voldoende kwaliteit. In dit whitepaper lichten we een aantal doelen toe van data governance indien ze relevant zijn voor data modellering

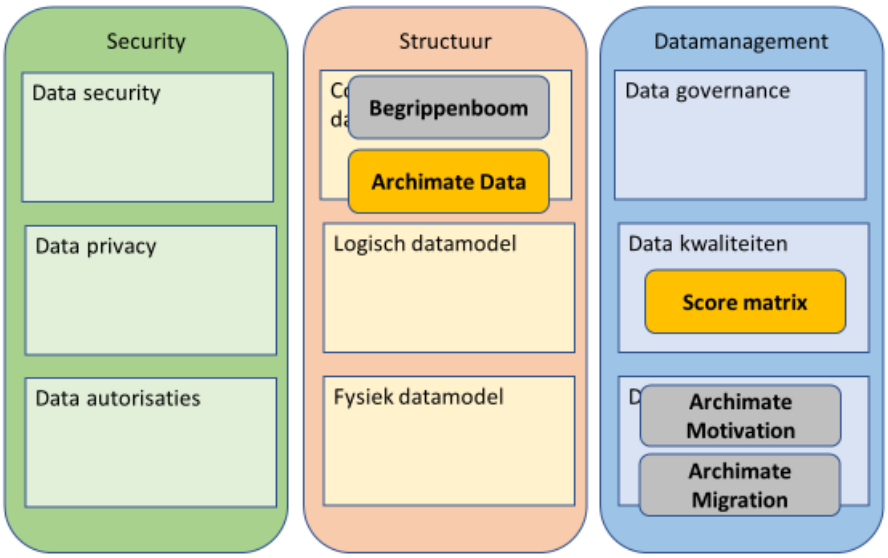
- **Verhogen data kwaliteit**, Bepalen van de data kwaliteit en het nemen van maatregelen die de kwaliteit van de data meetbaar verhogen.
- **Processen voor data kwaliteit,** inrichten van processen met ondersteuning van tools en methoden voor het continu verhogen van data kwaliteit voor de relevante datasets

DATA QUALITY EN DATA MODELLERING

Data quality en met name het bepalen van de kwaliteitsniveaus van datasets kan op eenvoudige wijze met data modellering worden uitgewerkt. Door dit te combineren met een aantal andere datamodelleerwijzen ontstaat een aantal weergaven cq viewpoints op de data die de huidige en gewenste kwaliteit kan bepalen van datasets. Daarnaast kunnen kwaliteit verhogende maatregelen gemodelleerd worden zodat de gap tussen huidige en gewenste situatie overbrugd kan worden

Notatiewijzen

Voor data modellering binnen data quality zijn een aantal notatiewijzen relevant. Een aantal is essentieel, en een aantal is ondersteunend. Onderstaande afbeelding geeft een beeld van de notatiewijzen die vervolgens kort worden toegelicht



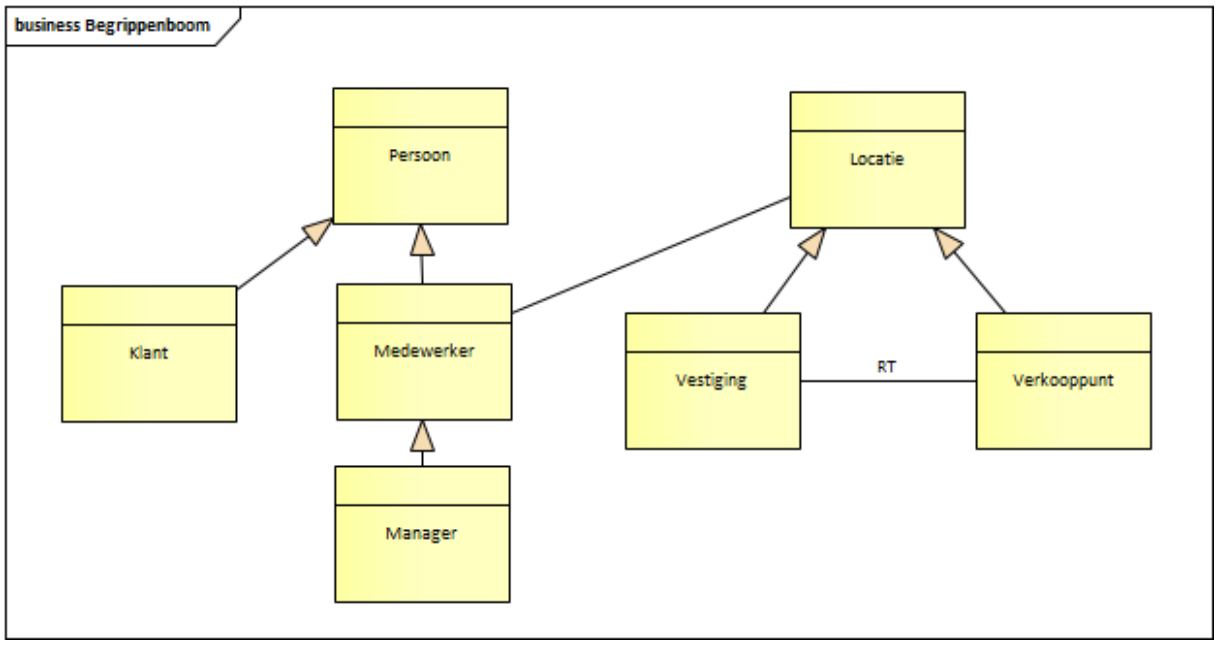
CONCEPTUEEL DATAMODEL

Het conceptueel datamodel is voor data quality een essentieel onderdeel dat zorgt voor de beschrijving welke data entiteiten cq data sets die relevant zijn binnen de organisatie. Dit is input voor het definiëren welk kwaliteitsniveau in de baseline geldt en wat het kwaliteitsniveau in de target inrichting dient te zijn

Voor het definiëren van de data kwaliteiten kan een relatief abstract model van de data entiteiten gebruikt worden. Details als de attributen en een detaillering van de associaties is veelal niet nodig. Vandaar dat kan volstaan met een korte definitie van de data entiteiten. Hierbij speelt het conceptuele data model en met name het ArchiMate datamodel een belangrijke rol.

Onderstaande afbeelding geeft een beeld van een ArchiMate datamodel. Meer informatie over de notatiewijze is te vinden via:

http://assistent.interactory.nl/cmsForm.aspx?formid=50027&webcontentid=249



Naast het uitwerken van de data entiteiten zal in een aantal organisaties de behoefte bestaan om naast de uitwerking van de entiteiten de definities uit te werken. Bijvoorbeeld wanneer door onduidelijkheid van de definities er interpretatieverschillen ontstaan. Dat is feitelijk een data kwaliteitsprobleem. Dan kan een begrippenboom met een bijbehorende lijst van definities als toevoeging aan het conceptuele datamodel meerwaarde brengen. Meer informatie over deze notatiewijzen is te vinden via:

http://assistent.interactory.nl/cmsForm.aspx?formid=50027&webcontentid=248

DATA QUALITY

Voor het modelleren van de data quality is in eerste instantie een overzicht van relevante data kwaliteiten van belang. Deze lijst kan men zelf opstellen, echter er zijn reeds een aantal standaard kwaliteitsindelingen aanwezig. Het model van DaMa is het meest compleet en wordt daarom veel ingezet.

Is er een lijst van kwaliteiten gedefinieerd dan kan deze gerelateerd worden aan de verschillende data entiteiten in de organisatie. Hiervoor zijn meerdere mogelijkheden, de meest eenvoudige, maar toch effectieve, aanpak is een scorematrix. Hierbij geef je in de matrix aan hoe een dataset scoort op een bepaalde kwaliteit. Dit wordt gedaan voor zowel de baseline als de target situatie. Datasets met een groot verschil tussen baseline en target kunnen vervolgens aangepakt worden met een aantal verbeterende maatregelen.

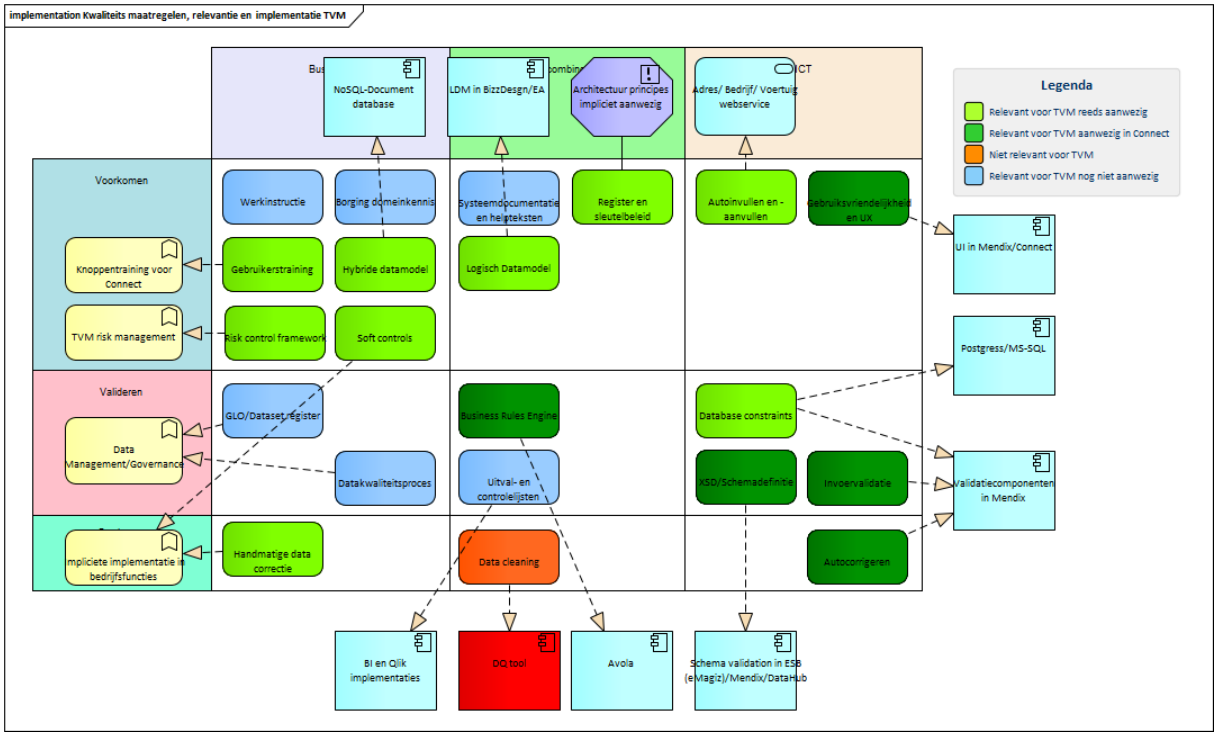
De score matrix een goed hulpmiddel. Score kan ingevuld worden met een getal tussen 0 en 10 of een ordinale indeling zoals Laag – Midden Hoog. Detailinformatie over de modelleervorm score matrix is te vinden via: http://assistent.interactory.nl/cmsForm.aspx?formid=50027&webcontentid=256

Target / Source	Accuraatheid	Actualiteit	Compleetheid	Consistentie	Precisie	Privacy	Redelijkheid	Referentiele integriteit	Tijdigheid	Uniekheid	Validiteit
Cursus	8	8	8	8	NaN		NaN	NaN	8	9	6
Docent	8	8	8	8	NaN	6	NaN	NaN	8	6	6
Training	8	8	8	8	NaN		NaN	NaN	8	9	6

DATA GEBRUIK

Voor de ICT afdeling, data eigenaar en de data steward is het handig als er inzage is welke kwaliteitsmaatregelen ingezet kunnen worden in de organisatie en welke relevant zijn binnen de eigen context. Dat is niet noodzakelijk maar zeker bij een complex landschap of een hoge volwassenheid van de organisatie rond data management zal aan een dergelijk kwaliteit – relevantie en maatregelmodel steeds meer behoefte ontstaan.

Data gebruik kan dan gemodelleerd worden met behulp van requirements maar ook op basis van workpackages die bepaalde kwaliteit verhogende maatregelen omvatten. De ArchiMate extensies Motivation en Migration hebben hier een notatiewijze voor. Onderstaande afbeelding geeft een voorbeeld van een koppeling tussen de bedrijfsprocessen en de entiteiten. Meer informatie over beide notatiewijzen is te vinden via:
http://assistent.interactory.nl/cmsForm.aspx?formid=50027&webcontentid=247



Kenmerken

Data quality komt bij steeds meer organisaties hoger op de prioriteitenlijst te staan. Problemen rond rapportages en compliance kunnen daaraan ten grondslag liggen. Maar ook het creëren van waarde uit data is een reden om te kijken naar data kwaliteit en te zoeken naar maatregelen die de kwaliteit verhogen

Data quality biedt vanuit data modelleringsperspectief een aantal interessante modelleerbehoeften, met name de combinatie van het conceptuele model met een lijst van generieke kwaliteiten en een bijbehorende score en daarnaast het modelleren van maatregelen is de kern in een data quality datamodel. Bij de introductie van data modellering van een data quality model zijn de volgende kenmerken relevant:

- Begin klein en start met een eenvoudig conceptueel datamodel zoals de begrippenboom
- Zorg voor de definities van de begrippen en leg verbanden met behulp van associaties
- Stel een lijst van data kwaliteiten op of gebruik een lijst vanuit één van de data standaarden
- Leg een score matrix aan om het de kwaliteitsniveaus van de baseline en de target te modelleren
- Werk eventueel de maatregelen voor data kwaliteit uit in ArchiMate modellen
- Verfijn het model en breidt het verder iteratief uit.

Producten

De producten voor een data governance vanuit data modelleringsperspectief zijn samengevat:

- Conceptueel datamodel
- Score matrix
- Model van maatregelen
- Modellen rond het gebruik van datasets in de organisatie

Tooling

Zoals reeds genoemd zijn er rond data quality meerdere producten te vinden, veelal als onderdeel van een data management suite. Gartner heeft hiervoor een aantal documenten opgesteld, een andere aardige ingang is: https://blog.panoply.io/12-data-management-tools

Ook generieke tooling kan ingezet worden. Inrichten op basis van Wiki's is een goede mogelijkheid, houdt bij de inrichting van een dergelijke omgeving direct rekening met het beheer en de governance van de entiteiten. Een dergelijke omgeving dient namelijk in sync te blijven lopen met de ontwikkelingen binnen de data kwaliteits inrichting en dat is geen eenvoudige opgave vanuit beheerperspectief.

Als laatste is het inzetten van generieke (enterprise) architectuurtooling te noemen. Een aantal architectuur tools hebben de mogelijkheid om meerdere modelleertalen met elkaar te combineren waardoor de (data) modelleerbehoefte voor data quality grotendeels kan worden afgedekt.

Evaluatie

Data quality is bij steeds meer organisaties een belangrijk werkveld. Inzetten van data quality kan veel redenen hebben, echter vrijwel altijd dient er een antwoord gevonden te worden op problemen rond de kwaliteit van data.

Binnen data quality speelt data modellering een belangrijke zo niet centrale rol. Met name het leggen van verbanden tussen de data entiteiten en een lijst van kwaliteiten inclusief een score is essentieel. In een vroeg stadium nadenken welke modelleervormen relevant zijn, hoe deze aan elkaar verbonden worden en hoe de stakeholders daarbij betrokken zijn ondersteunt de introductie van data quality.

In dit whitepaper hebben we een combinatie van modelleervormen beschreven die een (minimale) set is van notatiewijzen op basis waarvan data quality in organisaties gemodelleerd kunnen worden.

Over de auteur

Bert Dingemans is trainer op het vlak van data architectuur, data management en Big Data. Hij heeft een passie voor modelleren, modelleertools en het effectief inzetten van geautomatiseerde hulpmiddelen om modellen effectief in te zetten in de praktijk. Bert is te bereiken via bert@interactory.nl