De context van Meta Data

Juli 2023

Ir. Ing. Bert Dingemans

INHOUDSOPGAVE

INLEIDING

In dit White paper gaan we in op de context van meta data. Dat doen we vanuit de context van Meta Data Management. Meta data is een belangrijk kennisgebied van data management en daarmee relevant voor data gedreven organisaties.

We gaan daarbij in op verschillende gezichtspunten van meta data, wat is het precies, waarom is het belangrijk in data gedreven initiatieven en hoe kunnen we meta data met succes inzetten binnen organisaties. Dat doen we door eerst te kijken naar wat context is, we geven een aantal voorbeelden van meta data en gaan vervolgens in op diverse aspecten bij het verzamelen en inzetten van meta data

In de volgende hoofdstukken zien we dat meta data een nauwe relatie heeft met de andere kennisgebieden zoals uitgewerkt in de DaMa Body of Knowledge [DMBoK]. Zoals je zult zien is de grens tussen data en meta data niet altijd goed te trekken. Daarmee wordt het feit dat meta data met behulp van data management processen beheerd dient te worden.

WAT IS META DATA?

De definitie van meta data lijkt verrassend eenvoudig namelijk: Data over data. Echter wat betekent data over data? Een aantal meer gedetailleerde vragen over wat meta data is zijn:

- Wat is de structuur van meta data
- Wie maakt de meta data
- Is meta data relevant voor alle soorten data in een organisatie
- Zijn er meerdere soorten meta data?
- Kan er ook data zijn zonder meta data?
- Wie is eigenaar van de meta data?

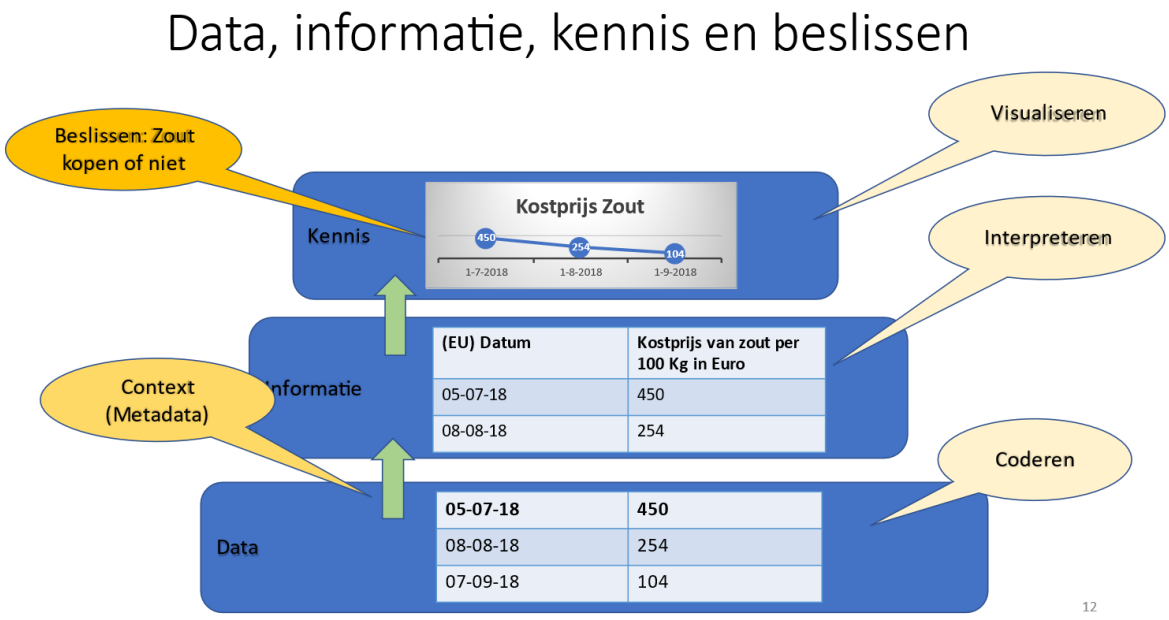
Allemaal vragen die aangeven dat meta data verschillende gezichten heeft. Daarmee is de term “data over data” wellicht wat te beperkt.

Een andere definitie is: Meta data geeft context aan data. Daarmee ga je dus proberen betekenis toe te kennen vanuit een bepaalde context of perspectief. In de volgende paragraaf geven we een voorbeeld van een veel gebruikte meta data classificatie.

Echter de meta data die context geeft aan een dataset of data entiteit wordt bepaald door de context van waaruit we naar de data kijken. De context bepaald daarmee welke meta data wordt vastgelegd over de data entiteit. Daarmee werkt context dus twee kanten op vanuit het gezichtspunt van de vraagsteller en vanuit het perspectief van de data entiteit met de bijbehorende meta data.

Bijvoorbeeld is de data over wie de eigenaar en steward zijn vanuit het perspectief van data governance? Of is een Beschikbaarheid, Integriteit, Vertrouwelijkheid en Privacy (BIVP) classificatie meta data vanuit het perspectief van data security en privacy? Beide vragen kunnen we positief beantwoorden. Dit is allemaal meta data over data vanuit een specifiek (data management) perspectief.

Meta data over data maakt het dus mogelijk om data om te zetten naar informatie. Deze informatie kan vervolgens geïnterpreteerd worden en op basis daarvan bouw je kennis op. Deze kennis maakt het dus mogelijk om uiteindelijke (gefundeerde) beslissingen te nemen. In de afbeelding wordt dit gepresenteerd.



Bron: [Masterclass Data Management]

Classificatie meta data

Er zijn verschillende soorten meta data classificaties. Hieronder een classificatie van [Dataversity]:

- **Beschrijvende metadata**: Dit type metadata wordt gebruikt voor ontdekking en identificatie. Het bevat beschrijvende elementen. Denk bijvoorbeeld aan de titel, auteur en trefwoorden.
- **Structurele metadata**: bevat beschrijvingen over data entiteiten. Het beschrijft de versie, relaties en andere kenmerken van digitaal materiaal.
- **Administratieve metadata**: presenteert informatie voor het beheer van een resource, zoals het resourcetype, machtigingen en hoe en wanneer de gegevens zijn gemaakt.
- **Referentie metadata**: Deze vorm van metadata gaat over de inhoud en kwaliteit van statistische gegevens.
- **Statistische metadata**: kan worden gebruikt om de processen te beschrijven die betrokken zijn bij het verzamelen, verwerken of produceren van statistische gegevens.
- **Juridische metadata**: het geeft informatie over de maker, de houder van het auteursrecht en openbare licenties.

Deze structuur geeft een naar mijn mening vrij willekeurige indeling. Kenmerkend is dat er verschillende gezichtspunten zijn die de meta data bepalen kijkend naar een bepaalde data

entiteit. In de volgende paragraaf kijken we naar de DMBoK die aangeeft dat er vele gezichtspunten zijn om meta data context te laten geven aan een data entiteit vanuit een DMBoK kennisgebied.

Meta data in de DMBoK

In de DMBoK is het kennisgebied Meta Data uitgewerkt en daar wordt het als een essentieel onderdeel van data management beschouwd. In het SIPOC model vanuit de DMBoK wordt een mooi overzicht gegeven van de kenmerken van Meta Data Management vanuit het data management perspectief.

Bron: [DMBoK]

VOORBEELDEN VAN META DATA

Meta data wordt vaak als abstract gezien. Dat komt mogelijk doordat meta data niet altijd direct zichtbaar is. Echter de meta data helpt ons wel om een inschatting te maken van de waarde die een data entiteit voor ons kan hebben. Het helpt ons om de meta data te interpreteren en daarmee uiteindelijk ook de data zelf te interpreteren. Hieronder twee eenvoudige voorbeelden van data in combinatie met meta data.

Foto meta data

Een digitale foto is opgebouwd uit data. De data van de foto maakt het mogelijk om deze te bekijken met een viewer. In een viewer heb je de mogelijkheid om ook enige meta data te zien over een foto. Hieronder een eigen vakantiefoto als voorbeeld van de weergave van de data in een viewer en aan de rechterkant enige meta data.

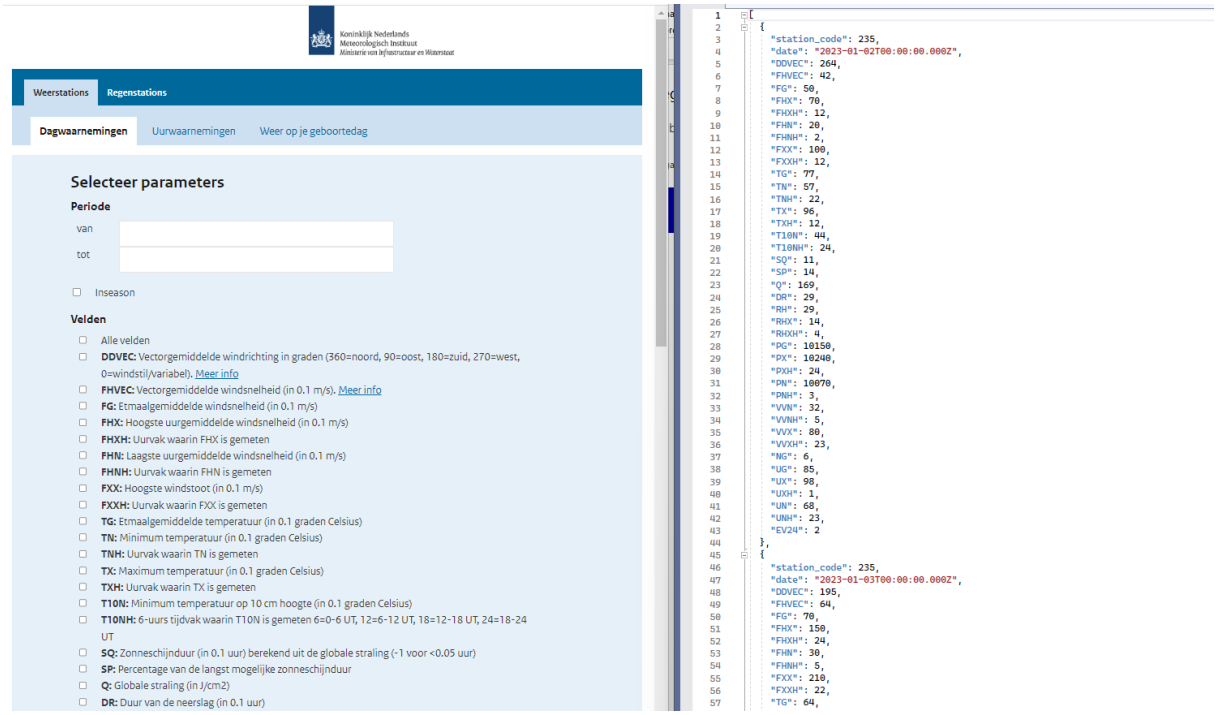
Interessant in deze is dat de camera waarmee deze foto gemaakt is meteen meta data heeft toegevoegd aan de foto. Zo zie je wanneer de foto gemaakt is, met welk apparaat en wat informatie over het formaat en de locatie van de foto op mijn bestandssysteem.

Dit is meta data die al wel wat informatie geeft over de data van deze foto. Maar vanuit een andere context mist er wel wat meta data. Bijvoorbeeld waar is deze foto gemaakt, door wie is die foto gemaakt en met welke reden. Daarmee zie je dat de bruikbaarheid van aanwezige meta data bepaald wordt door de context van de vraag die je stelt over de data of welke meta data mist om de gestelde vragen te kunnen beantwoorden.

Klimaat meta data

Op de website van het KNMI kun je interactief klimaatgegevens opvragen. Daarmee zie je een mooie combinatie van een data set of data entiteit in combinatie met meta data. In de afbeelding hieronder zie je links een formulier waarin je een aantal meta data elementen interactief kunt selecteren. Rechts zie je een voorbeeld van de inhoud van de dataset opgebouwd op basis van de meta data zoals gedefinieerd door het KNMI. In de meta data is in

detail beschreven wat de betekenis is van de verschillende attributen is die in de dataset zijn opgenomen. De meta data kan nog verder geïnterpreteerd worden door via hyperlinks extra meta data te verkrijgen over hoe de klimaatdata verzameld wordt en wat de kwaliteitsniveaus zijn van de data in de set. Zonder de meta data kun je de data in de data set onmogelijk interpreteren wat de data zonder meta data feitelijk waardeloos maakt. [https://daggegevens.knmi.nl/klimatologie/daggegevens].



HERKOMST VAN META DATA

In de twee voorbeelden in de vorige paragraaf is een combinatie van data en meta data visueel inzichtelijk gemaakt. Bij het gebruik van meta data is het vergaren van meta data rond de data een belangrijke activiteit. De herkomst van meta data is in te delen in een aantal soorten. Daarbij kun je bij de classificatie ook kijken naar de effort die het kost om deze meta data te verzamelen en de te verwachte datakwaliteit en datamodellering te bepalen.

Bron geleverde meta data

Door de bron van de data wordt de meta data meegeleverd met de data. Soms is dat door in de data zelf de meta data op te nemen. Bij de werkwijze met bestanden en bestandsuitwisseling wordt de werkwijze van meta data insluiten vaak toegepast. In het voorbeeld van de foto heeft de camera een stukje meta data aan de data toegevoegd zoals de datum en het type camera etc. Deze werkwijze heeft vanzelfsprekend in zich dat slechts een beperkte set aan meta data kenmerken opgenomen kan worden.

Daarnaast is het ook mogelijk dat er naast een data een verwijzing is naar de meta data waar de beschrijving over de karakteristieken in de meta te raadplegen is. Veel toegepast bijvoorbeeld bij koppelvlakken, open data sets en web APIs. Het voorbeeld van de KNMI website rond de weerdata is hier een mooi voorbeeld van.

Afgeleide meta data

Bij het werken met data sets is een deel van de meta data af te leiden uit de data zelf. Zoals in het voorbeeld met de klimaatdata kun je meta data afleiden. Bijvoorbeeld in deze data kun je afleiden in de structuur van de data wat het aantal rijen en kolommen in een bepaalde dataset is, welke kolommen erin voorkomen. Ook interessant zijn een aantal diepgaandere analyses van de inhoud van de data. Bijvoorbeeld de spreiding van de data in de kolommen, de verdeling, domeinen in de data, afwijkingen in de structuur en statistische analyses.

Deze afgeleide meta data heeft als voordeel dat het grotendeels geautomatiseerd verzameld kan worden voor en na de datavergaring. Echter veelal is deze meta data niet compleet voor inzet binnen een bepaalde context. Zoals in het voorbeeld van de klimaatdata, je kunt de kolomnamen afleiden op basis van de dataset, maar de betekenis van de inhoud van deze kolommen zal nietszeggend zijn zonder een beschrijving van de gebruikte afkortingen voor de kolomnamen. Daarvoor is toch echt de webpagina met de meta data noodzakelijk.

Impliciete meta data

Impliciete meta data is een interessante categorie. Het maakt namelijk gebruik van kennis die impliciet aanwezig is rond deze data door ervaring en de opbouw van kennis. Als ik deze combinatie van tekens geef 6713 KA dan is bij de meeste Nederlandse gebruikers de impliciete kennis aanwezig dat dit een Nederlandse postcode is. Het voorbeeld van de foto kan ook impliciete bevatten, iemand die al eerder in Stralsund aan de Oostzee is geweest zal weten dat de locatie van die foto in die Hanzestad is.

Impliciete meta data is relatief eenvoudig te vergaren omdat het direct aanwezig kan zijn. Echter de kwaliteit van de aanwezige metadata is niet altijd correct. Bijvoorbeeld de

cijfercombinatie 13045 is data waarvan je niet weet of het een Duitse of Franse postcode is. Dus impliciete meta data kan interessant zijn maar heeft vanuit datakwaliteiten perspectief een aantal risico's en vraagt daarom inzet van vergaren van meta data door nadere analyse en inzet van menselijke kennis. Mogelijk dat op termijn nieuwe kunstmatige intelligentie hierbij ingezet kan worden om hiermee de afgeleide meta data interpretatie te verbeteren.

Vergaarde meta data

Dit is de meta data die vergaard moet worden rond data entiteiten, tijdens het verzamelen, transformeren, gebruiken en managen van de data. Bij deze data moet men dus actief deze meta data vergaren met name voor de diverse contexten rond meta data. Denk bijvoorbeeld aan de data rond data management processen zoals data governance en data kwaliteiten.

Dit is data die actief vergaard moet worden met beperkte geautomatiseerde ondersteuning om deze meta data te vergaren. Bijvoorbeeld wie is de data eigenaar of -steward is meta data die actief vergaard moet worden, veelal door inzet van mensen die deze data vergaren en registreren. Ook bijvoorbeeld de data rond transformatie vanuit de bron naar het uiteindelijk gebruik is een vorm van vergaarde meta data. Soms kan dit deels vergaard worden met geautomatiseerde hulpmiddelen, echter het heeft bijna altijd ook een handmatig deel in zich.

In de voorgaande paragrafen hebben we gekeken naar vier soorten meta data herkomst. Hieronder een tabel met een eenvoudige verdere categorisering van deze vormen van meta data met daarbij een indeling naar de kenmerken van deze meta data herkomst. Dit kan vervolgens gebruikt worden om een inschatting te maken hoe de meta data herkomst inzetbaar is in de eigen context.

Herkomst	Data vergaring	Data kwaliteit	Data model	Inspanning
Bron data	Geautomatiseerd	Afhankelijk van de bron	Vooraf bepaald	Variabel
Afgeleide meta data	Geautomatiseerd	Variabel	Deels te bepalen	Laag
Impliciete meta data	Geautomatiseerd en handmatig	Variabel	Deels te bepalen	Laag tot Middel
Vergaren meta data	Handmatig en deels geautomatiseerd	Definieerbaar	Vrij te definiëren	Hoog

META DATA GERELATEERDE ASPECTEN

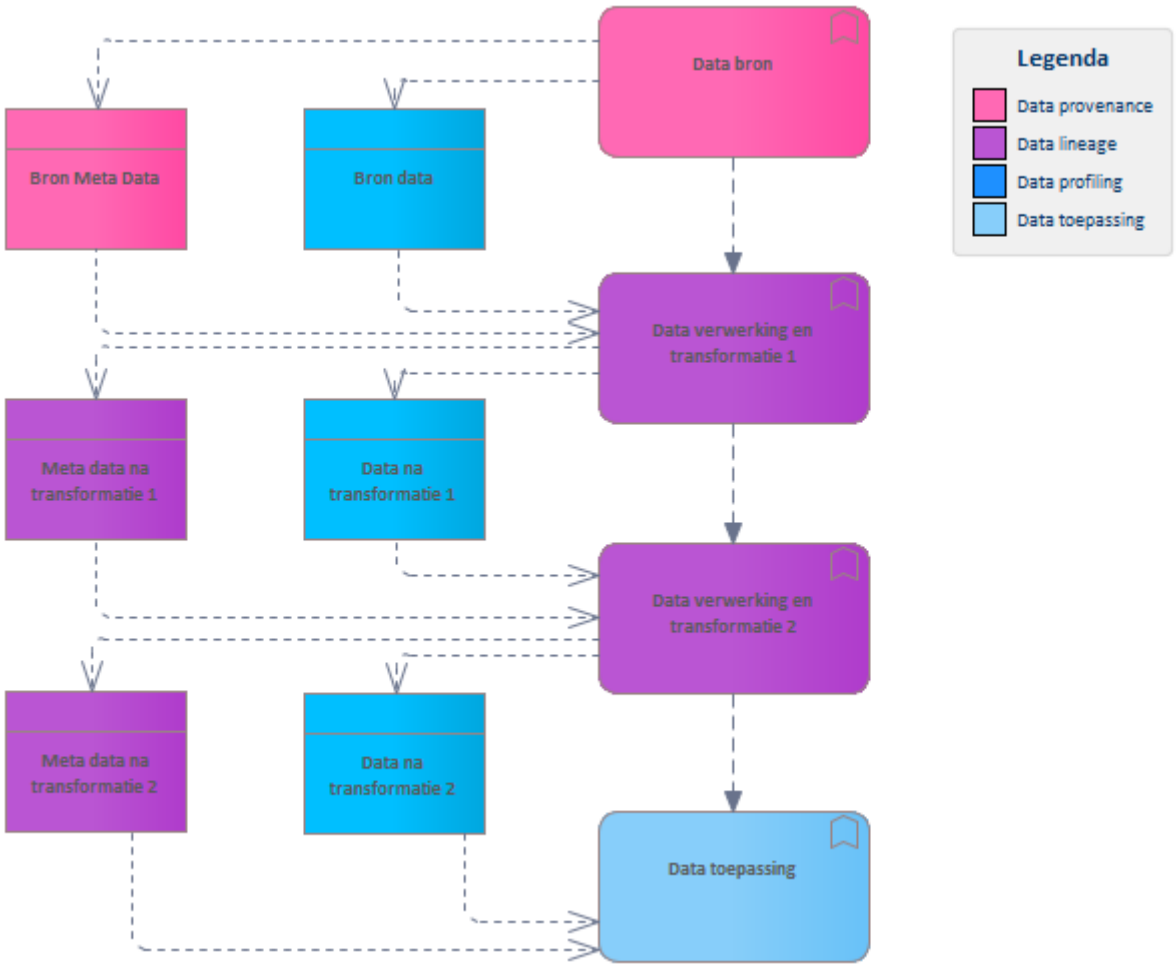
Tot nu toe hebben we voornamelijk gekeken naar de bronnen van data en de structuur die de data. Echter in de meeste gevallen zal de data afkomstig uit interne- en externe bronnen getransformeerd moeten worden naar een structuur wat de data voor onze toepassing geschikt maakt. Deze transformaties kan meerdere redenen hebben maar een aantal kenmerkende redenen zijn

- De structuur van de data uit de bron heeft een andere structuur dan de gewenste structuur bij de toepassing
- De kwaliteit van de data uit de bron is te laag en dient door transformatie naar een voor de toepassing gewenst niveau gebracht te worden.
- De data is afkomstig uit meerdere bronnen en dient aan elkaar gerelateerd te worden op basis van identificerende kenmerken zoals sleutels
- Vanuit security of privacy perspectief dient de data uit de bron geanonimiseerd of gemaskeerd te worden voordat het ingezet kan worden in de toepassing
- De data uit de bron heeft een formaat dat ongeschikt is voor analyse en visualisatie ten behoeve van de toepassing.

In de voorgaande hoofdstukken hebben we gezien dat meta data als dat door de bron geproduceerd wordt een mooi startpunt is voor het vergaren van meta data. Echter in de vervolgstappen gaan we de brondata transformeren naar andere vormen voor de uiteindelijke data toepassing. In de volgende afbeelding zie je een transformatie van de eerdere foto getransformeerd naar een pentekening weergave. Echter als we de meta data opvragen van deze afbeelding dan is de meta data niet veranderd. Dat betekent dan ook dat in de situatie van de foto we zelf meta data zullen moeten vergaren om inzichtelijk te maken dat we een data transformatie hebben toegepast van een foto naar een pentekening van de originele data uit de bron, de digitale camera.

Bij de activiteiten tussen de databron en de datatoepassing kunnen dus een aantal transformaties zitten. Deze transformaties grijpen in op de data in de verschillende databewerkingen die plaatsvinden. In de volgende afbeelding tonen we dit in een eenvoudig ArchiMate diagram. De data komt als een data object uit een bron in combinatie met meta data, vervolgens worden er transformatie functies uitgevoerd op de data en wordt er over deze transformatie meta data geproduceerd. Hiermee ontstaat een stapsgewijze transformatie van de data en een stapsgewijze vergaring van meta data die uiteindelijk beiden ingezet kunnen worden in de data toepassing.

Voor dataverwerking en het transformeren van data in relatie tot meta data zijn een drietal kenmerkende activiteiten te noemen. Deze activiteiten worden in de volgende paragrafen beschreven.



Data provenance

Provenance wordt gedefinieerd als een record dat de mensen, instellingen, entiteiten en activiteiten beschrijft die betrokken zijn bij het produceren, beïnvloeden of leveren van data. Met name de provenance of herkomst van data is cruciaal om te beslissen of deze data te vertrouwen is voor gebruik in de toepassing, hoe deze moet worden geïntegreerd met andere databronnen en hoe de afzenders moeten worden vermeld bij hergebruik. In een open en inclusieve omgeving zoals het web, waar gebruikers data vinden die vaak tegenstrijdig of twijfelachtig is, kan herkomst die gebruikers helpen om een vertrouwensoordeel van de herkomst te vormen.

Vanauit meta data is dit een belangrijk activiteit omdat we reeds eerder hebben gezien dat datakwaliteiten in het gehele traject een rol spelen en uitgangspunt vormen voor het selecteren van een bepaalde databron en bepalend is voor de transformaties die plaats dienen te vinden.

Data lineage

Data lineage omvat de oorsprong van de gegevens, welke bewerkingen erop plaatsvinden en waar ze in de loop van de tijd naartoe gaan. Data lineage geeft zichtbaarheid en vereenvoudigt tegelijkertijd de mogelijkheid om fouten terug te voeren naar de hoofdoorzaak in een gegevensverwerkings en – analyseproces na de databron.

Het maakt het ook mogelijk om specifieke delen of invoer van de gegevensstroom opnieuw af te spelen voor stapsgewijze foutopsporing of het regenereren van verloren uitvoer. Databasesystemen gebruiken dergelijke informatie, data-provenance genoemd, om vergelijkbare validatie- en debugging-uitdagingen aan te pakken.

Herkomst van gegevens verwijst naar bronnen uit systemen en processen die van invloed zijn op de gegevens die van belang zijn, en biedt een historisch overzicht van de gegevens en de oorsprong ervan. Het gegenereerde bewijs ondersteunt forensische activiteiten zoals analyse van gegevensafhankelijkheid, detectie en herstel van fouten/compromis, auditing en nalevingsanalyse. "

Data lineage kan visueel worden weergegeven om de gegevensstroom/beweging tussen de bron naar de bestemming (toepassing) te ontdekken via verschillende veranderingen tijdens de transformaties, hoe de gegevens onderweg worden getransformeerd, hoe de weergave en parameters veranderen en hoe de gegevens splitsen of convergeren na elke bewerkingsstap. Een eenvoudige weergave van de gegevenslijn kan worden weergegeven met transformatiefunctionaliteiten en verbindingen tussen deze transformatiepunten. [Wikipedia]

Data profiling

Dataprofilering verwijst naar de analyse van data voor gebruik in een toepassing zoals een datawarehouse om de structuur, inhoud, relaties en afleidingsregels van de gegevens te verklaren. Profilering helpt niet alleen om anomalieën te begrijpen en de datakwaliteit te beoordelen, maar ook om meta data van organisaties te ontdekken, te registreren en te beoordelen. Het resultaat van de analyse wordt gebruikt om de geschiktheid van de potentiële databronnen te bepalen, meestal als basis voor een vroege go/no-go-beslissing, en ook om problemen te identificeren voor later data lineage ontwerpen.

Gegevensprofilering maakt gebruik van methoden van beschrijvende statistiek zoals minimum, maximum, gemiddelde, modus, percentiel, standaarddeviatie, frequentie, variatie, aggregaten zoals aantal en som, en aanvullende metadata-informatie verkregen tijdens gegevensprofilering, zoals gegevenstype, lengte, discrete waarden , uniciteit, voorkomen van null-waarden, typische tekenreekspatronen en abstracte typeherkenning. De meta data kunnen vervolgens worden gebruikt om problemen op te sporen, zoals illegale waarden, spelfouten, ontbrekende waarden, weergave van verschillende waarden en duplicaten. Allemaal gericht op de verhoging van de kwaliteit van de data.

In voorgaande paragrafen hebben we een aantal meta data aspecten en activiteiten beschreven die een bijdrage leveren aan meta data tussen de data bron en de data toepassing. In die tabel hieronder een overzicht welke activiteiten gerelateerd zijn aan welke herkomsttypen van meta data. Op basis van onderstaande tabel kun je kijken welke activiteit toegepast kan/moet worden wil je in over een bepaalde context meta data verzamelen en welke combinatie van activiteiten de beste datakwaliteitsverhoging tot gevolg zal hebben. Dit inclusief hoeveel effort je moet steken in het vergaren van kwalitatief voldoende meta data tussen data bron en -toepassing.

Herkomst	Data provencance	Data Lineage	Data profiling
Bron data	X		
Afgeleide meta data	X	X	
Impliciete meta data	X	X	X
Vergaren meta data		X	X

META DATA PATRONEN

In het vorige hoofdstuk zijn we ingegaan op de activiteiten die een rol spelen bij meta data rond datavergaring. In dit hoofdstuk wordt ingegaan op meta data patronen. Patronen zijn voor veel data architecten een passie, waaronder ondergetekende.

Data patronen zijn generieke oplossingen voor frequent terugkerende meta data problemen. Voor dataverwerking en -vergaring is dit kenmerkend rond meta data. Dataverwerking en -transformatie is van zichzelf al een uitdaging. De behoefte om op adequate wijze ook de meta data te verzamelen en te registreren is daarmee een extra complicerende factor. In onderstaande paragrafen worden daarom een aantal kenmerkende meta data patronen beschreven.

Meta Data Register

Het meta data register is een data gedreven toepassing met een aantal bijzondere kenmerken met name op het vlak van gebruikerswensen zoals zoeken, filteren, relaties leggen en visualisaties.

Probleem

Meta data wordt in veel toepassingen gebruikt en dient dan ook raadpleegbaar te zijn voor meerdere soorten stakeholders. Allemaal hebben zij hun eigen kijk (of context) op meta data van databron tot -toepassing.

Context

De context van meta is relatief breed. Kenmerkend is dat data management hierin een essentiële rol vervuld. De rollen dataeigenaar, -steward en -architect zijn verantwoordelijk voor het bepalen van de requirements van alle stakeholders.

Op basis daarvan wordt een toepassing als meta data register aangeschaft, geconfigureerd of zelf ontwikkeld. Daarmee is de context van de meta data dus de gehele organisatie in de breedste zin van het woord. Namelijk ook stakeholders van buiten de eigen organisatie kunnen gebruik maken van het door de organisatie aangeboden data uit het meta data register.

Oplossing

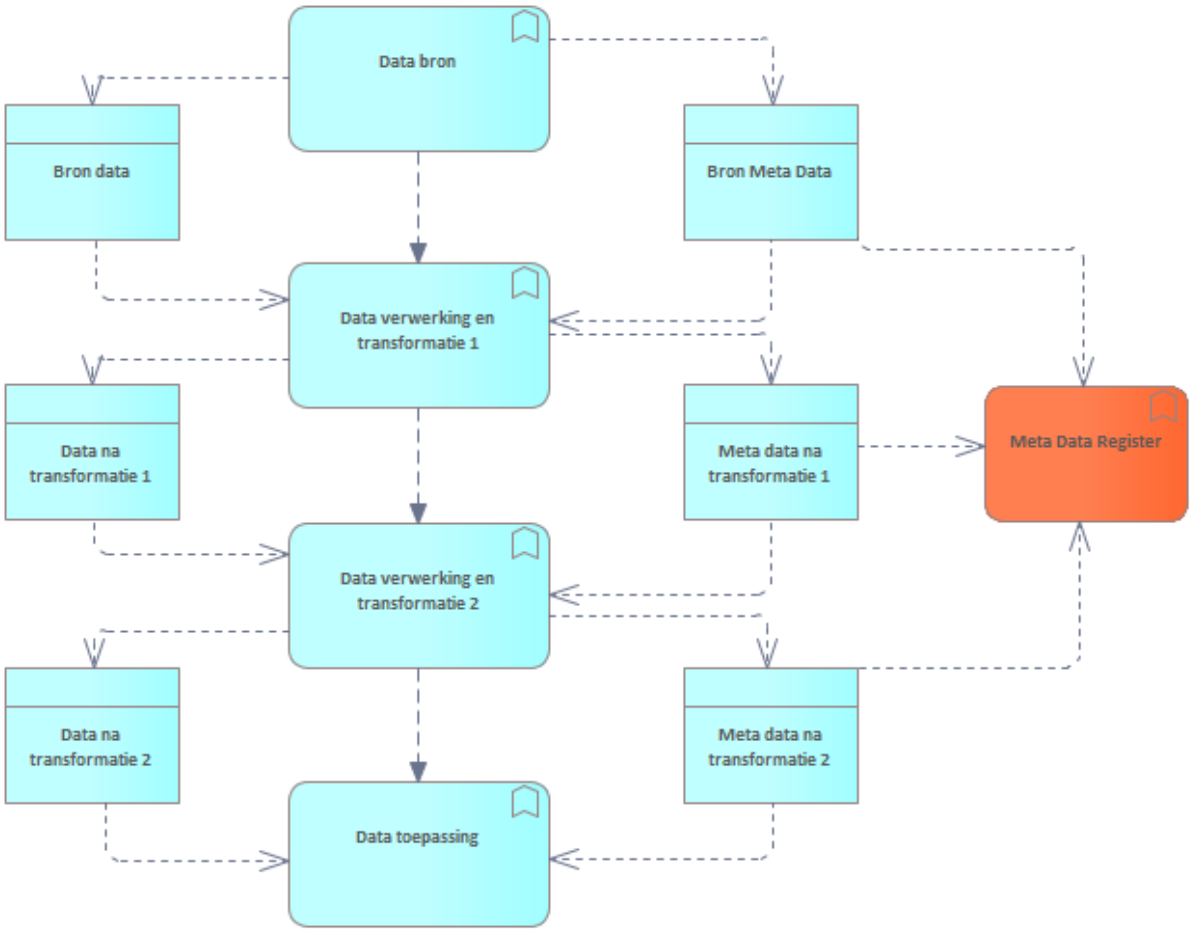
Zoals al aangegeven in de vorige paragraaf is de oplossing in meerdere vormen mogelijk, namelijk standaard software, configuratie van een bestaand modelleerpakket of eigen maatwerk voor een meta data register.

Kenmerkend in de oplossing is dat het onderliggende model van een meta data register gebaseerd is op een kenmerkend datamodel dat is opgebouwd uit:

- Entiteiten
- Relaties tussen entiteiten
- Attributen

Bij de inrichting van het datamodel van een meta data register zal op basis van iedere context een specifiek datamodel geconfigureerd worden voor de diverse stakeholders. Er is een ISO standaard aanwezig over meta data. Deze ISO standaard is voor een aantal standaard software als uitgangspunt genomen, andere producten hebben een eigen inrichting gemaakt. In een whitepaper van de Meta Data werkgroep is hiervoor een model uitgewerkt op basis van een combinatie van standaard modelleertalen.

Onderstaande ArchiMate diagram toont hoe vanuit de stappen tussen de databronnen en de data transformaties de meta data overgebracht wordt naar het meta data register. Het meta data register is dan weer bron voor de verschillende soorten gebruikers (stakeholders) van het meta data register.



Rationale

Meta data is veelomvattend en kent vele gezichten. Echter een register waar de meta data te raadplegen is biedt mogelijkheden om de vele verschillende verzoeken van de stakeholders te standaardiseren in de implementatie van een meta data register.

Meta Data Insertion

Dit patroon gaat in op het verzamelen van meta data bij de dataverwerking, bij voorkeur daadwerkelijk tijdens de dataverwerking zelf.

Probleem

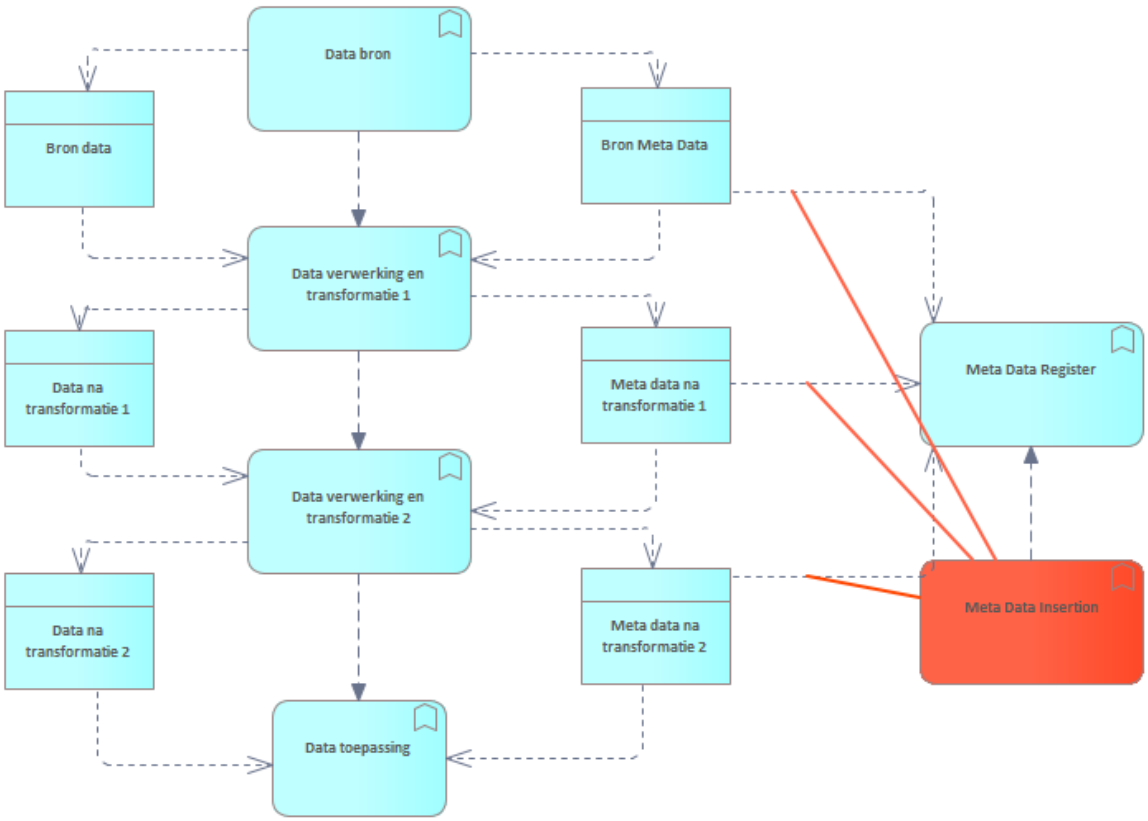
Data transformatie en -verwerking tussen de data bron en -toepassing kent vele vormen en implementaties. Dat betekent feitelijk dat ook tussen de dataverwerking functionaliteiten en het meta data register als data toepassing een data transformatie implementatie ingericht moet worden naar het meta datamodel in het meta data register. Gezien de vele vormen van data verwerking is dit daardoor een complexe implementatie.

Context

De context van meta data insertion zit voornamelijk binnen de (technische) implementatie van datatransformatie toepassingen. Soms is dit als implementatie beschikbaar in data transformatie tools (ETL) of in (big) data platformen en data lakes.

Echter veel organisaties kennen een divers data transformatie landschap van data transformaties en data verwerking waardoor meta data insertion bij de inrichting van een dergelijke omgeving speciale aandacht vraagt van onder andere de data architect.

Oplossing



De oplossing is een patroon dat tijdens de data transformatie en verwerking de relevante aspecten van de daadwerkelijke transformaties verwerkt tot meta data en deze vervolgens naar het meta data register verstuurd. Binnen de meta data insertion worden ook weer een aantal data patronen gebruikt zoals data egress en event- of trigger gebaseerde dataverwerking.

Rationale

Meta data insertion vindt plaats tijdens de dataverwerking en niet achteraf dat is een belangrijk voordeel ten opzichte van het patroon in de volgende paragraaf meta data harvesting. Met name ten behoeve van de data kwaliteiten tijdigheid en actualiteit is de instante verwerking een belangrijk kenmerk van meta data insertion.

Meta Data Harvesting

Probleem

Meta data harvesting wordt gedaan als dataverwerking reeds heeft plaatsgevonden in het verleden zonder dat men toen meta data over de transformatie verzameld heeft. In die situatie dient data harvesting met terugwerkende kracht dataverwerkingsalgoritmen te ontdekken en te analyseren.

Data harvasting is vooral relevant in situaties waar het ontstaan van de huidige data architectuur evolutionair ontstaan is zonder rekening te houden met eisen en requirements die vanuit meta data management gesteld worden.

Het analyseren van de programmatuur die zorgdragen voor de data transformaties kan complex zijn. Zeker in situaties waar weinig gebruik gemaakt is van standaardisatie van transformaties, meerdere data verwerkingstools zijn gebruikt of waar de geschreven software door meerdere professionals ontwikkeld zijn, waarbij logging e.d. niet is ontwikkeld, kan meta data harvesting een uitdaging zijn.

Context

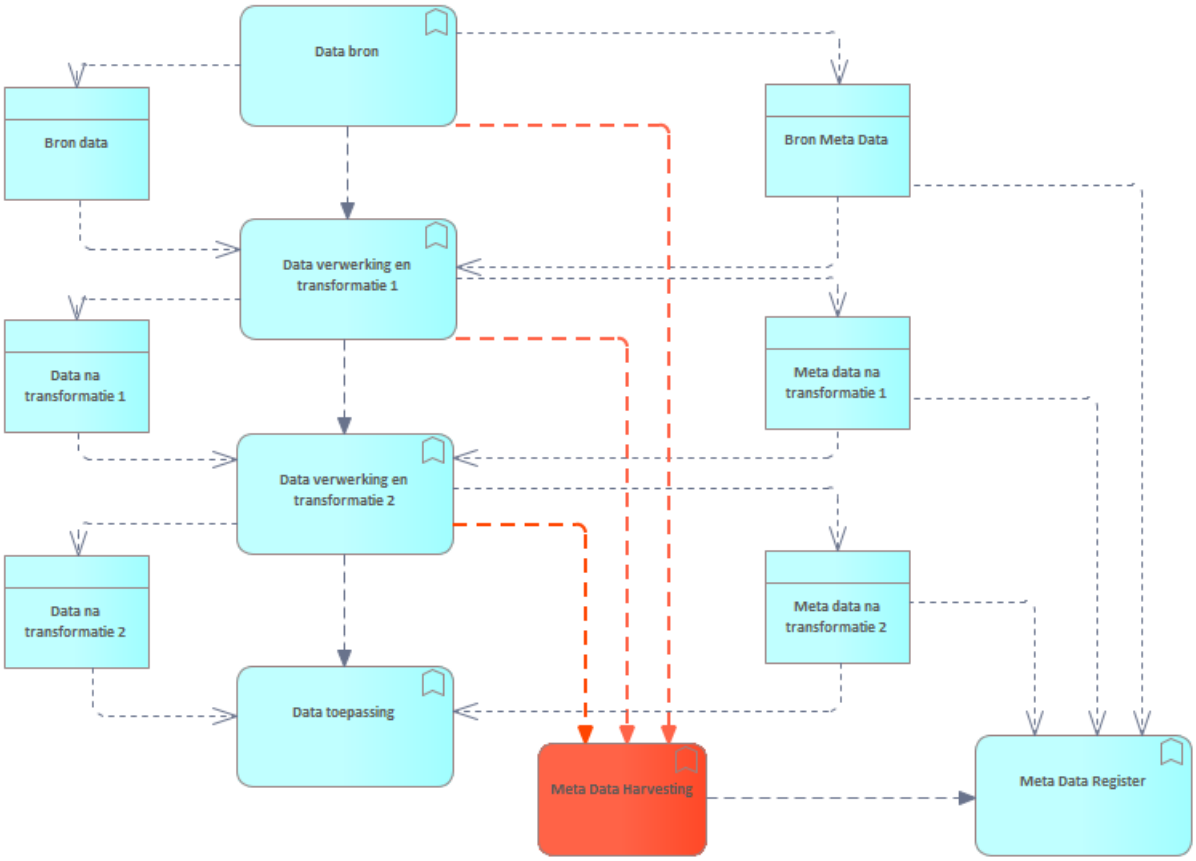
Meta Data Harvesting is vooral relevant in de context van Business Intelligence, Master - en Referentie Data en data integratie relevant. Men dient met terugwerkende kracht meta data te verzamelen over de dataverwerking binnen deze werkvelden. In deze werkvelden zijn de kansen op succesvolle implementaties groot omdat hier de automatiseringsgraad van de dataverwerking hoog zal zijn gezien het repeterende en gestandaardiseerde karakter van de toepassingen die deze vormen van dataverwerking implementeren.

Oplossing

De oplossingen voor meta data harvesting bestaan uit geautomatiseerde hulpmiddelen die in staat zijn de reeds uitgevoerde data transformaties tussen de data bron en de -toepassing op te sporen en te transformeren naar een model voor meta data. Deze transformatie naar meta data voor het meta data register zal veelal bestaan uit een combinatie van fysieke- en mogelijk logische data modellen van de structuur en de tussenresultaten van data modellen tussen de transformaties te analyseren.

De oplossing kan bestaan uit traditionele data harvesting technieken, zoals het lezen dan fysieke data modellen uit relationele databases of het gebruik van gestandaardiseerde transformatietechnieken zoals XSLT.

Echter in complexe situaties kunnen meer geavanceerde harvesting technieken noodzakelijk zijn. Denk hierbij aan kunstmatige intelligentie gebaseerd op het zoeken naar patronen in de implementatie van programmatuur het analyseren van logs van de gebruikte data transformatie toepassingen en andere systemen die laag gestructureerde data produceren over de uitgevoerde data transformatie.



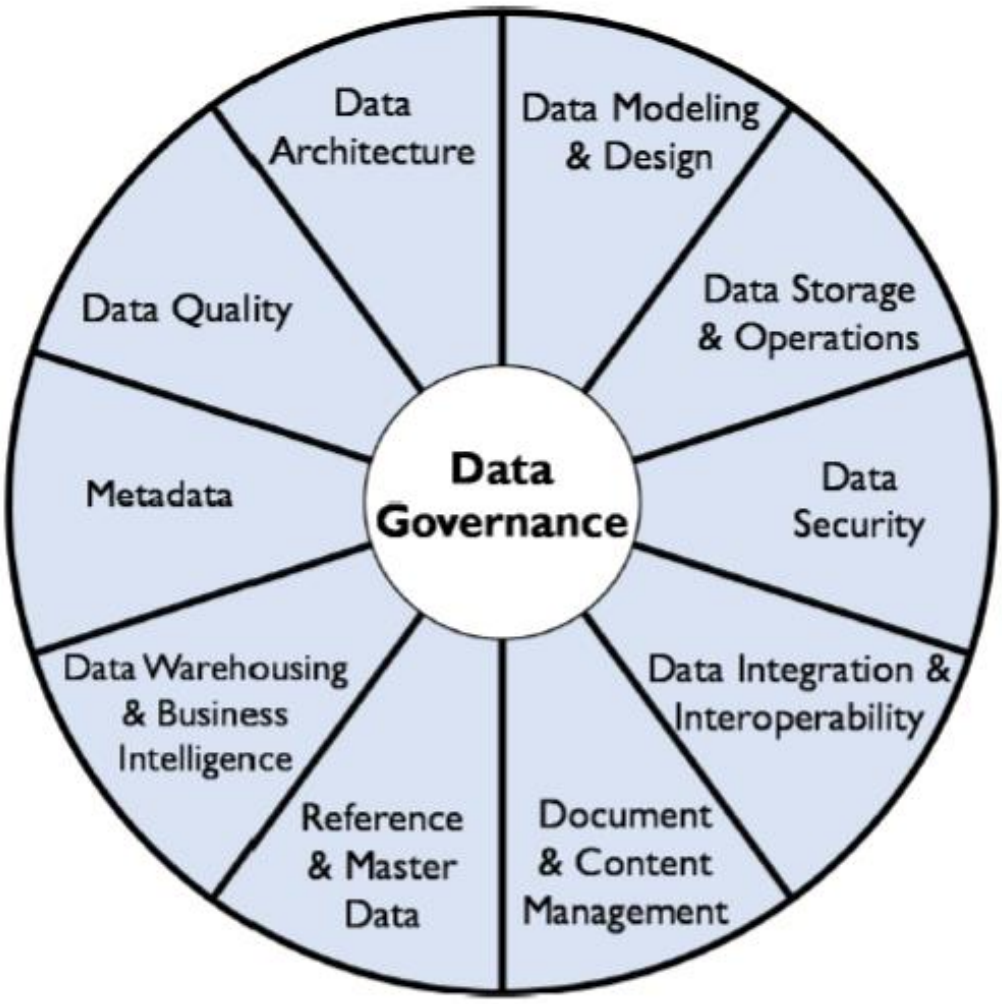
In bovenstaande ArchiMate diagram wordt de meta data harvesting functie toegevoegd tussen de bestaande data verwerking en transformatie stappen en het meta data register. Deze meta data harvesting functie is in dit diagram niet verder uitgewerkt, echter deze functie zal uitgesplitst kunnen worden in meerdere harvesting functies die gecombineerd tot het gewenste resultaat kunnen leiden.

Rationale

Meta data management rond de data bronnen, transformaties en het gebruik is een belangrijk aspect om de context van meta data rond deze activiteiten in kaart te brengen. Hierdoor neemt de volwassenheid van meta data toe wat indirect een positief effect zal hebben op de andere kennisgebieden binnen Data Management. Meta Data Harvesting maakt het met terugwerkende kracht mogelijk meta data te vergaren.

RELATIE TOT ANDERE DATA MANAGEMENT KENNISGEBIEDEN

Meta Data Management heeft een link met alle andere kennisgebieden binnen het DMBoK raamwerk. In alle kennisgebieden wordt een context gemaakt van de data die binnen dit kennisgebied verzameld wordt en daarmee is impliciet de relatie met meta data gelegd. In dit whitepaper zullen we echter de relatie met de belangrijkste kennisgebieden binnen DaMa kort toelichten.



Bron: Dama.org

Data governance

Data governance gaat over de autoriteit van data en daarmee over de autoriteit van de meta data. Er is daardoor een nauwe relatie tussen data governance en meta data. Vanuit data governance dient beleid geformuleerd te worden wat de doelen zijn rond meta data management. Echter meta data levert naar data governance de context van de data in de organisatie.

Impliciet wordt de relatie tussen data governance en meta data management gelegd, zeker in organisaties waar data management nog in de kinderschoenen staat. Men zoekt naar de context van data in de verschillende kennisgebieden maar doet dat weinig gestructureerd. Vaak gebaseerd op een reactieve en issue gedreven werkwijze bij het introduceren data management.

Het structureren van de activiteiten binnen meta data management kan een organisatie ondersteunen om grip te krijgen op de activiteiten binnen data management. Hiermee wordt een proactieve werkwijze gezocht. Meta Data wordt dan feitelijk een administratieve toepassing voor data governance en de andere kennisgebieden waarmee bij data management issues de meta data reeds aanwezig is.

Data architectuur

Meta data management en data architectuur hebben een nauwe relatie. Zoals reeds gezegd is meta data van belang in de context van alle kennisgebieden. Data architectuur zijn de activiteiten om kaders te stellen aan verandering rond data in een organisatie. Deze architectuurkaders dienen ervoor te zorgen dan door veranderingen er geen wildgroei ontstaat aan inrichtingen van datavergaring, -transformatie en -gebruik. Kaders worden veelal door de architecten in samenwerking met de data governance rollen uitgewerkt in de vorm van data architectuur principes.

Deze principes hebben veelal een directe relatie met meta data. Denk aan de uitwerking van principes rond de data governance rollen over de data of de structuur van de data maar ook de kwaliteitseisen die aan data gesteld worden. De data architect zal vervolgens deze principes inzetten in projecten en programma's waarin verandering geïntroduceerd wordt door op basis van de organisatie principes de implicaties van de verandering in relatie tot meta data te beschrijven. Een implicatie is daarbij kenmerkend namelijk: welke meta data dient verzameld te worden over de verandering die geïntroduceerd wordt in de organisatie betreffende data management.

In het vorige hoofdstuk zijn we ingegaan op datapatronen. Patronen zijn belangrijke concepten en hulpmiddelen voor de architect. Hiermee kan er zowel een beschrijvende- en voorschrijvende beschrijving gegeven worden vanuit de architect naar de verandering die plaatsvindt. Er zijn slechts een paar meta data patronen uitgewerkt, de data architect zal een collectie van meta data patronen opstellen waarmee er hergebruik en standaardisatie geïntroduceerd wordt in de organisatie gericht op meta data.

Data modelleren

Meta data en data modelleren zijn nauw met elkaar verbonden. Datamodellen zijn feitelijk zelf meta data over de structuur van de data. De meta data zal vaak beschreven worden in de vorm van een data model. Bij gestructureerde data worden modelleertechnieken gebruikt voor de drielaagsmodellering: conceptueel, logisch en fysiek. De modelleertechnieken toegepast zijn zelf gebaseerd op een eigen meta model. Modelleertechnieken zoals ArchiMate en UML kennen een uitgewerkt metamodel dat inzetbaar is voor meta data modellering. Meta data wordt daarmee uitgedrukt in de vorm van een datamodel.

Vanuit meta data is het van belang om de context van de verschillende kennisgebieden goed te kennen. Daarmee wordt feitelijk de mate van uitwerking en detail van de verzamelde meta data van deze kennisgebieden in een meta datamodel beschreven. Hiervoor kunnen gestandaardiseerde datamodellen ingezet worden maar ook eigen data modellen voor meta data worden uitgewerkt.

Data kwaliteiten

Binnen veel organisaties is de kwaliteit van de data een uitdaging en daarmee een reden om met data management aan de slag te gaan. Hierbij is er een relatie tussen meta data en data kwaliteiten. Wat is de context van de data kwaliteit. Welke kwaliteitsdimensies zijn relevant in de organisatie. Wat zijn de huidige en de gewenste kwaliteitsniveaus bij het gebruik van de data, maar ook wat is de kwaliteit aan de databronnen en welke data transformaties hebben effect op de data kwaliteiten, zowel positief als negatief.

Voor datakwaliteit zal dan ook een meta data model uitgewerkt dienen te worden. Hierbij is er direct een verbinding te leggen naar data governance, -architectuur en -modelleren voor het realiseren van voldoende datakwaliteit. Meta data is daarmee nauw verweven met deze kennisgebieden.

Data integratie en Data warehousing

Data integratie en data warehousing worden in dit whitepaper gecombineerd beschreven vanuit het perspectief van meta data. Beide kennisgebieden zijn nauw met elkaar verbonden en worden gecombineerd toegepast. Beiden richten zich op het ontsluiten van data bronnen, het transformeren van data naar een uitwerking gericht op een data toepassing. Of deze toepassing gericht is op het krijgen van inzicht uit data of het overbrengen van data van het ene naar de andere informatiesysteem, bedrijfsproces of ketenpartner geeft aan dat de relatie met meta data relevant is.

In de vorige hoofdstukken is gebleken dat het verzamelen van meta data over aspecten als bronnen, transformatie en gebruik niet altijd goed is ingericht. Hierdoor ontstaat een probleem rond de meta data die verzameld wordt over deze activiteiten, als deze meta data sowieso al verzameld wordt.

Hiermee wordt ook hier de context, en dus meta data, van belang vanuit het perspectief van data governance, -architectuur en – kwaliteiten van belang. De beschreven patronen in het vorige hoofdstuk zijn dan ook een startpunt voor het introduceren van data management in het

algemeen en de meta data die verzameld wordt over data integratie en data warehousing is daarvoor een belangrijke bron van meta data.

Overige kennisgebieden

De overige kennisgebieden van het DMBoK data management raamwerk worden in dit whitepaper niet verder behandeld. Betekent niet dat er geen relatie is met meta data. Van alle kennisgebieden is context belangrijk. Context wordt uitgewerkt door meta data over deze kennisgebieden te verzamelen op een gestructureerde wijze gebruik maken van de vele hulpmiddelen, zoals patronen, principes en modelleerwijzen die aanwezig zijn rond de context van de diverse kennisgebieden.

TOOLING & METHODEN

Meta data management is een complex werkveld dat daarbij nauw verwezen is met de andere data gerelateerde kennisgebieden. Daarnaast is het gevolg van het verzamelen van meta data dat er omvangrijke datasets ontstaan met daarin meta data over de data toegepast binnen de organisatie.

Veelal reden om voor meta data management te zoeken naar tools en methoden die de activiteiten en de bijbehorende registratie van meta data (geautomatiseerd) ondersteunen. Methoden en technieken zijn bestaande raamwerken zoals de DMBoK, open standaarden en generieke meta data modellen die ingezet worden bij het inrichten van meta data management.

Gezien de omvang van de registratie en de meta datasets wordt de inzet van meta data tools onontbeerlijk. Bij meta data tools zijn meerdere scenario's mogelijk:

- Er zijn diverse standaard meta data management tools verkrijgbaar. Deze tools hebben een generieke uitwerking van de meta data management processen. Daarnaast zijn veelal een aantal hierboven genoemde patronen geïmplementeerd zodat de omvangrijke registratie van meta data geautomatiseerd verzameld en beheerd kan worden. Voorbeelden van dergelijke meta data tools zijn Collibra, Axon en Talend.
- Werkt een organisatie reeds met een generiek modelleertool dan kan meta data registratie in een dergelijk generiek tool worden uitgewerkt, Bijvoorbeeld indien een combinatie van meerdere modelleertalen ondersteund wordt, waaronder ArchiMate, UML en ER en de matrix modellen, kan een werkwijze zijn om een meta data register te implementeren. Voorbeelden zijn Sparx Enterprise Architect, Visual Paradigm of Blue Dolphin.
- In organisatie waar een gedisciplineerde inrichting aanwezig is van kennismanagement en samenwerkingsomgevingen kan een meta data register op deze wijze geïmplementeerd worden. Men sluit dan aan op de reeds aanwezige uitgewerkte kennis in dergelijke platformen op basis van een vooraf uitgewerkt metamodel voor de meta data. Voorbeelden van dergelijke toepassingen zijn Wiki's (Confluence) of SharePoint.

EVALUATIE

In dit whitepaper is ingegaan op de context van meta data. Zoals kenmerkend bij meta data is context het centrale begrip is. Echter de context heeft zowel een vraag- als aanbod gedreven verschijning. Dat maakt dat meta data de verschillende contexten, requirements en toepassingen samenbrengt.

Naast een beschrijving van de herkomst van meta data is een link gelegd naar de databonnen, -verwerking en -transformatie en de datatoepassing vanuit het perspectief beschreven. Deze herkomst wordt vervolgens uitgewerkt in een aantal veelgebruikte data gerelateerde activiteiten zoals data lineage, - provenance en profiling. Deze activiteiten worden

ondersteund door patronen waarvan wij er drie hebben beschreven namelijk meta data register, meta data insertion en meta data harvesting.

In het laatste hoofdstuk leggen we een relatie tussen meta data management en de andere kennisgebieden van data management zoals uitgewerkt in het DMBoK. Tot slot noemen we kort een aantal mogelijke methoden en tools om een inrichting van een meta data register te introduceren.

Graag vermeld ik de reeds aanwezige artikelen, whitepapers en voorbeelden van meta data zoals opgesteld door de meta data werkgroep van DaMa-NL [MetaData] en beschikbaar op de website over metadata.

VERWIJZINGEN

[DaMa]: https://www.dama.org/cpages/home

[DaMa-NL]: https://dama-nl.org

[Data-Docent]: https://data-docent.nl/

[Dataversity]: https://dataversity.net/

[DMBoK]: https://technicspub.com/dmbok/

[MetaData]: https://dama-nl.org/werkgroepen/werkgroep-metadata/

[Wikipedia]: https://www.wikipedia.org/

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