Boek van Chisholm echte aanrader voor wie grip wil hebben op business

Antwoord op vragen rond semantiek en databeheer

Stijn Christiaens

Malcolm Chisholm schreef een boek over het belang van het opstellen en beheren van definities, los van specifieke applicaties. Ook de theorie rond definities komt aan bod. Hier volgt een review van Chisholm's "Definitions in Information Management: A Guide to the Fundamental Semantic Metadata".

Malcolm Chisholm geeft in de inleiding het hoofddoel van zijn boek duidelijk aan: "to appreciate definitions in the context of information management". Hij voert een heleboel gerespecteerde experts in het veld op, die allemaal bevestigen dat duidelijke definities de hoeksteen van informatiebeheer vormen. De volgende quote vat het boek mooi samen, en geeft aan waarom je het zou moeten kopen en lezen:
"This book looks at what definitions are, the different types of definitions, how they are used, how high-quality definitions can be constructed, what problems can exist in definitions, *how to manage definitions, and particularly at how to justify work on definitions to executive management*" (cursief zelf toegevoegd).

Vanaf het begin van het boek hecht Chisholm veel belang aan de motivatie voor definities: ze zijn essentieel voor het goed begrijpen van business concepten (klant, inschrijving, rekening) in data (in spreadsheets, rapporten, databases), ze disambigueren termen (klant in de context van sales versus boekhouding), ze verklaren of je bezig bent met metadata of met data (datum van update), ze verbinden instanties met hun concepten, ze zijn nodig voor consistentie bij afleidingen en berekeningen (als alternatief voor de *reverse engineering* van een algoritme in een black box), ze helpen bij het vergelijken van concepten in data mapping en ze maken controle mogelijk voor *drift* in de datavelden. Chisholm identificeert dat definities een grote inspanning vragen bij source data analysis (SDA), en stelt dat bestaande definities helpen om herhaaloefeningen in analyse te vermijden en data-integratie te vergemakkelijken.

Uitvoerige studie

Chisholm is zich er absoluut bewust van dat semantiek (als de studie in betekenis) niet nieuw is. Zijn boek vormt dan ook een uitvoerige studie van de bestaande literatuur rond definities, die hij helder relateert aan de problemen van databeheer. Er hangt momenteel een hype rond het Semantic Web[1] of Linked Data[2] als een nieuwe manier om met data om te gaan. In plaats van in te gaan op de hype, geeft Chisholm antwoorden op de vragen rond semantiek waarmee data professionals over de hele wereld dagelijks geconfronteerd worden: "wat betekent X?" (wanneer ze het tegenkomen in een rapport, database, meeting, XML). Chisholm geeft duidelijk het verschil aan tussen een symbool, een term, een concept, een instantie en hoe ze bij elkaar passen. Afbeelding 4.2 uit het boek beeldt dit eenvoudig en duidelijk af. Het boek informeert de lezer over het belang van het opstellen en beheren van definities, los van specifieke applicaties. Dit is precies hoe het moet gebeuren. Gedurende de evolutie van informatiesystemen is er altijd een trend naar losse koppeling om complexiteit beheersbaar te houden: databases beheren data voor verschillende applicaties, de three tier architectuur, webservices als meer granulaire blokken van business logica, processen en regels als *first class citizens* die herbruikt worden in verschillende applicaties. Hetzelfde geldt voor definities: wanneer de organisatie een definitie voor klant heeft, wil je dat die gealigneerd is en gebruikt wordt in alle applicaties en systemen, eerder dan dat ze stilletjes uitdooft op een blad papier.
Chisholm identificeert welke data- en metadata-objecten definities nodig hebben: databasetabellen en -kolommen, applicatieschermen en schermlabels, applicatierapporten, rapportlabels en interfacebestanden, business concepten, entiteiten, hun attributen en relaties.

Het boek levert ook theorie rond definities: echte definities ten opzichte van nominale definities. Echte definities leveren uitleg rond het wezen van het concept, terwijl nominale definities de betekenis van een woord of term beschrijven. Daarnaast beschrijft Chisholm ook een typologie van definities (essentiële, onderscheidende, oorzaak/verband, incidenteel, ostensief, stipulatief, legislatief en ondefinieerbaar). Hij wijst op het belang van het juiste type op de juiste plaats, bij voorkeur vanuit een strategisch perspectief te bepalen.

Aangezien definities een essentieel onderdeel vormen van elk initiatief rond datakwaliteit, is de juiste kwaliteit van definities van groot belang. Chisholm wijdt dan ook een volledig hoofdstuk aan de verschillende kwaliteitsaspecten voor definities. Vanuit een methodologisch perspectief kunnen deze kwaliteitsaspecten

gemakkelijk ingebouwd worden in een eigen organisatorische aanpak voor kwalitatieve definities. Op deze manier zorg je dat de hoeksteen van het databeheer stevig in elkaar zit. Het is immers niet zo nuttig om te investeren in wat Bob Seiner 'cheeseburger definities' noemt[3].

Chisholm begrijpt goed dat het niet altijd vanzelfsprekend is om definities te bouwen en te beheren; "It is not possible to expect that all definitions will be complete at the outset. We come to know something initially and then gradually get to know it better".

Het bouwen en beheren van een laag van definities betekent door een soort van 'vaagheidstrechter' heen bewegen: je moet altijd beginnen vanuit heel gewone termen in al hun vaagheid, maar stapsgewijs daal je in de trechter af (bijvoorbeeld door het toevoegen van beschrijvingen, voorbeelden, nota's) tot de definities opgesplitst zijn in hun concepten, feiten en regels. Chisholm raakt daarna ook de onderwerpen precisie (het niveau van detail tot waar de data betrouwbaar zijn) en accuraatheid aan (het niveau waarop de data voorstellen wat ze horen voor te stellen).

Scope en context

Het boek reserveert twee hoofdstukken voor twee belangrijke aspecten van definities: scope en context. Scope gaat over het kiezen van de grenzen van je speelterrein, en Chisholm somt de volgende op: populaties en subpopulaties, subklassen van meer generieke concepten en verzamelingen van verschillende dingen. Wat context betreft, beschrijft hij de volgende gebieden van een organisatie die een context kunnen leveren: dochterondernemingen, lines of business, horizontale businessfuncties, geografische gebieden en toepassingen. Zijn perspectief op het vaak gebruikte voorbeeld van klant is even duidelijk als confronterend: "What is the purpose of an enterprise-wide definition of 'customer'?"

In een later hoofdstuk heeft Chisholm het over *governance*, een onderwerp dat vaak gezien wordt als invasief, terwijl dat zeker niet zo hoeft te zijn. Volgens het boek gaat governance over het instellen van rollen, verantwoordelijkheden en rechten met betrekking tot data. Hier wijst Chisholm terug op het belang van openheid in tegenstelling tot het verbergen van definities in diepe velden in een duistere legacy applicatie. Definities moeten afzonderlijk bewaard, beheerd en gepubliceerd worden, beschikbaar en ontsloten voor iedereen in de organisatie, met de juiste verantwoordelijkheden (aangeduid als 'trusteeship' in het boek). Chisholm behandelt verschillende aspecten van governance beknopt in het hoofdstuk: van validatie en verificatie tot monitoring, evaluatie en metrieken.

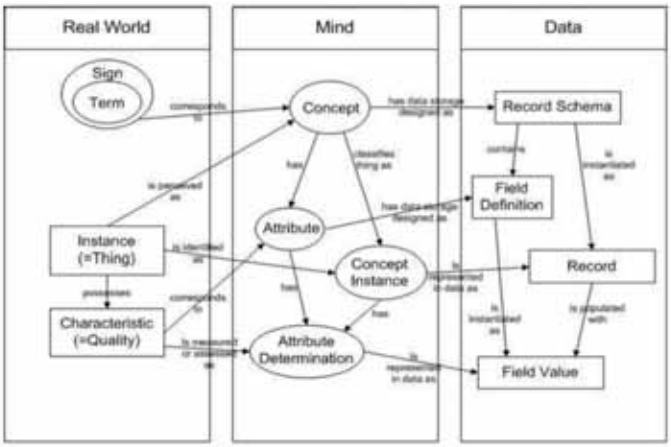

Figure 4.2: *Preliminary Illustration of Data and What It Represents*

De laatste hoofdstukken in het boek focussen op metadata voor definities (mooi verwijzend naar Dublin Core[4]) en geven bruikbare conclusies, bijvoorbeeld:

- executive management kan de kosten voor het maken en beheren van definities begrijpen als de juiste aanpak gevolgd wordt;
- het maken en beheren van definities is een proces dat niet past binnen de scope van een enkel tijdgebonden project.

Chisholm volgt zijn eigen advies, en sluit het boek af met een gedetailleerde woordenlijst, definities van Dublin Core, een aanzet tot een metamodel en een niet-triviaal voorbeeld om je te helpen bij het definiëren van je eigen business.

Het boek heeft een hanteerbaar aantal pagina's, waarmee het de focus houdt op het kernonderwerp, dat het goed en in detail uitlegt. Ik kijk alvast uit naar de volgende extra's op de website van het boek (www.data-definition.com):

- Meer uitleg over beschikbare standaarden om definities te beheren, zoals de Semantics of Business Vocabulary and Rules (SBVR) standaard van de Object Management Group (OMG). De inhoud van dit boek lijkt sterk in lijn met de standaard;
- Een overzicht van de beschikbare technologie om met definities en hun governance om te gaan;
- Een overzicht van methodologieën om een laag van definities te verkrijgen en te onderhouden, zowel binnen als buiten de grenzen van de organisatie.

Globaal gezien raad ik dit boek zeker aan voor alle professionals die een beter begrip willen krijgen van hun eigen business, of ze nu starten met een nieuw Business Intelligence programma (of al ver in een datawarehouse met onduidelijke data zitten), of dat ze aan het uitzoeken zijn hoe te voldoen aan complexe regelgeving, of gewoonweg nieuwe applicaties of databases bouwen.

Om meer te leren over het boek of om het te kopen; kijk op www.data-definition.com.

Noten

1. http://en.wikipedia.org/wiki/Semantic_Web
2. http://en.wikipedia.org/wiki/Linked_Data
3. Definition of cheeseburger: a cheeseburger is a burger with cheese.
4. http://www.dublincore.org/

Stijn Christiaens is mede-oprichter en operationeel directeur van Collibra (www.collibra.com), een enterprise softwarebedrijf dat business semantics toevoegt aan SOA-, EAI- en data-integratieprojecten.