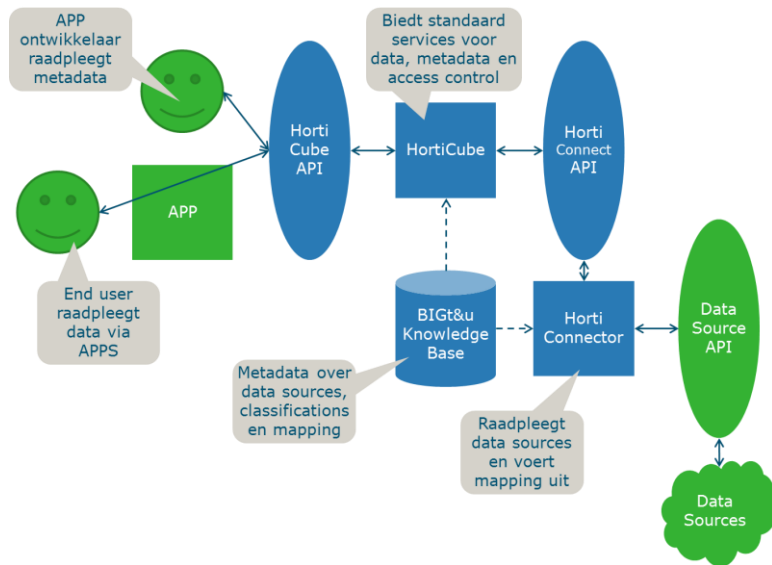


## Domeinmodel van de metadata in de HortiCube

Versie 6, 23 juni 2016

### Inleiding

De HortiCube levert via gestandaardiseerde interfaces gestandaardiseerde data aan applicaties. De functionaliteit van de HortiCube is gebaseerd op metadata over beschikbare data, dat wil zeggen data die beschrijven welke data via de HortiCube opgevraagd kunnen worden, hoe de data sets gestructureerd zijn, welke classificaties en meeteenheden gebruikt worden in een data set, hoe de data uit de bron opgevraagd kunnen worden, aan welke voorwaarden voldaan moet worden om de data te mogen gebruiken, welke mappings beschikbaar zijn om de data op andere classificaties af te beelden, zodat data uit



Figuur 1. Overzicht van de functie van HortiCube en HortiConnector

verschillende databronnen aan elkaar gekoppeld kunnen worden. De metadata zijn opgeslagen in een kennisbank, die ook kan worden geraadpleegd door ontwikkelaars van applicaties om de juiste queries te kunnen formuleren. De HortiCube roept voor het ophalen van data uit de bronnen de HortiConnector aan, die de metadata in de kennisbank gebruikt voor het raadplegen van de bronnen en het uitvoeren van mappings. Figuur 1 geeft een overzicht van deze functies.

Een belangrijk onderdeel van de metadata is de informatie over classificaties. Classificaties zijn opsommingen van entiteiten en klassen van entiteiten waarover data beschikbaar zijn, zoals landen en productindelingen. In verschillende databronnen worden vaak verschillende classificaties voor hetzelfde begrip gebruikt. Het land Nederland kan bijvoorbeeld worden aangeduid met de code 0003, NL of NLD. De HortiConnector kan mappings uitvoeren op basis van de metadata. Onder mapping verstaan we het vervangen van de codering uit de databron door een gewenste classificatie. Daarmee wordt het voor ontwikkelaars eenvoudig gemaakt om data uit verschillende bronnen aan elkaar te koppelen.

Het domeinmodel beschrijft welke metadata beschikbaar is en hoe die gestructureerd is. Dat is enerzijds van belang voor de ontwikkeling van de HortiCube zelf, anderzijds voor applicatieontwikkelaars die de HortiCube gebruiken om data te verkrijgen en te koppelen. De metadata worden beschikbaar gesteld als Linked Open Data en via het API van de HortiCube.

Bijlage 1 bevat een overzicht van het domeinmodel van metadata in de HortiCube. In het model zijn de volgende deelgebieden te onderscheiden:

1. Data Sets
2. Classifications
3. Semantics
4. Scopes of Data Set
5. Conversions and Mappings

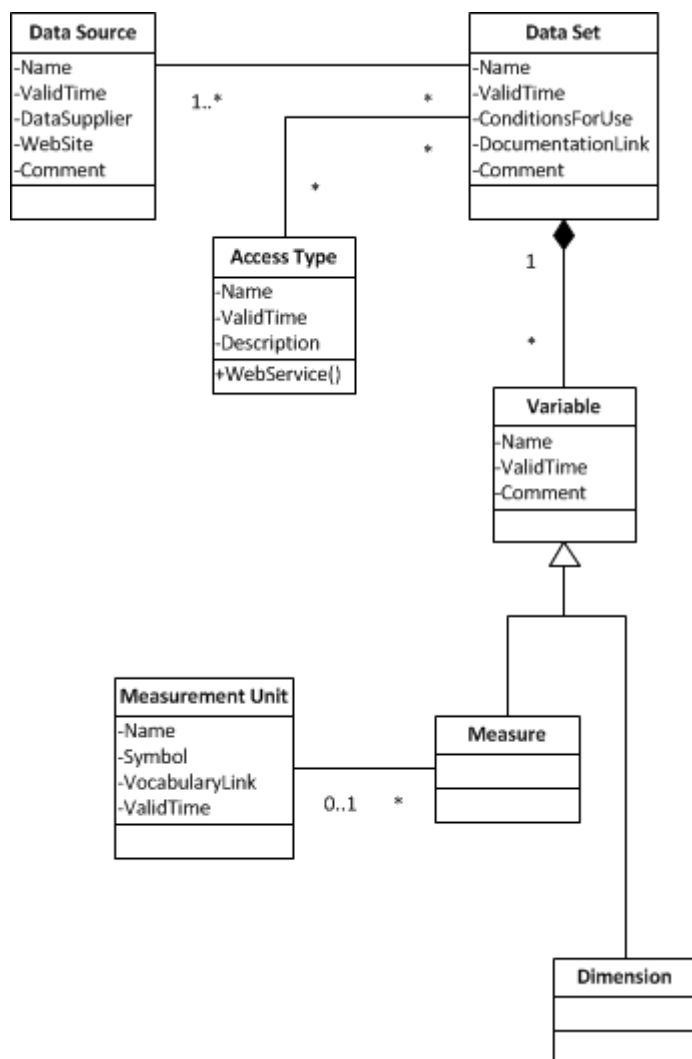
In onderstaande paragrafen worden volgens deze indeling views op het domeinmodel besproken. Daarbij komen alle classes uit het domeinmodel aan de orde. Voor de leesbaarheid zijn de modellen getekend als UML class models. Het is evenwel de bedoeling om de metadata als Linked Open Data in een RDF-formatted ontology beschikbaar te stellen.

## 1. Data Sets view

De Data Sets view beschrijft data sets die via de HortiCube benaderd kunnen worden. De structuur is weergegeven in Figuur 2.

Een **Data Set** beschrijft de relatie tussen een aantal variabelen. Een Data Set wordt geïdentificeerd door een Name, bijvoorbeeld "Oogstprognose van appels en peren". De geldigheid van de beschrijving kan beperkt zijn tot data over een bepaald tijdvak (ValidTime). Voor verschillende, niet overlappende tijdvakken kunnen verschillende versies van de beschrijving gelden. In de beschrijving kunnen een tekstuele beschrijving van de ConditionsForUse, een URI van de plaats waar documentatie te vinden is (DocumentationLink) en Comment (tekstuele aanwijzingen voor gebruik en interpretatie van de data) opgenomen zijn.

Een Data Set verwijst naar een **Data Source** (of meerdere data sources, als de informatie daaruit is samengevoegd in de Data Set). Een Data Source wordt geïdentificeerd door een Name. Verder bevat de beschrijving van een Data Source de naam van de Data Supplier (de organisatie die de data levert), een link naar een WebSite waar informatie te vinden is, en eventueel Comment van de HortiCube-beheerder. Er kunnen in de loop van de tijd meerdere versies van de beschrijving bestaan, bijvoorbeeld als de link naar de WebSite verandert of Comment wordt toegevoegd. ValidTime verwijst in dit geval naar de periode van geldigheid van de informatie over de Data Source, niet naar de geldigheid van de Data zelf. Een voorbeeld van een Data Source is "European Apple & Pear Crop Forecast August 2014", van de DataSupplier "World Apple and Pear Association (WAPA)", WebSite: <http://www.wapa-association.org/>.



Een Data Set bevat een verwijzing naar een **Access Type**. Dit is een technische beschrijving van de manier waarop de HortiConnector de data opvraagt.

De **Variables** in een Data Set hebben een Name, kunnen slechts voor bepaalde periodes beschikbaar zijn met een bepaalde semantiek, en zijn getypeerd als **Measures of Dimensions**.

Een **Measure** is een numerieke variable, geïdentificeerd door een Name, bijvoorbeeld "Production". Dit is de naam waaronder de Variable door software uit de Data Set kan worden opgevraagd.

Met een Measure kan verwijzen naar een **Measurement Unit**. Voor "Production" kan dat bijvoorbeeld "Ton" zijn, met symbool "tn". Een Measurement Unit heeft een VocabularyLink, de URI van de plaats op het World Wide Web waar de Unit als LOD gedefinieerd is, bijvoorbeeld als item van een vocabulary of concept in een ontology.

Een **Dimension** kwalificeert de Measures in de Data Set. In een dataset met de Measure "Production" kunnen bijvoorbeeld de Dimensions "Country", "Variety" en "Year" voorkomen.

Een Data Set kan gezien worden als een matrix met Variables als kolommen.

Figuur 2. Data Sets view

## 2. Classifications view

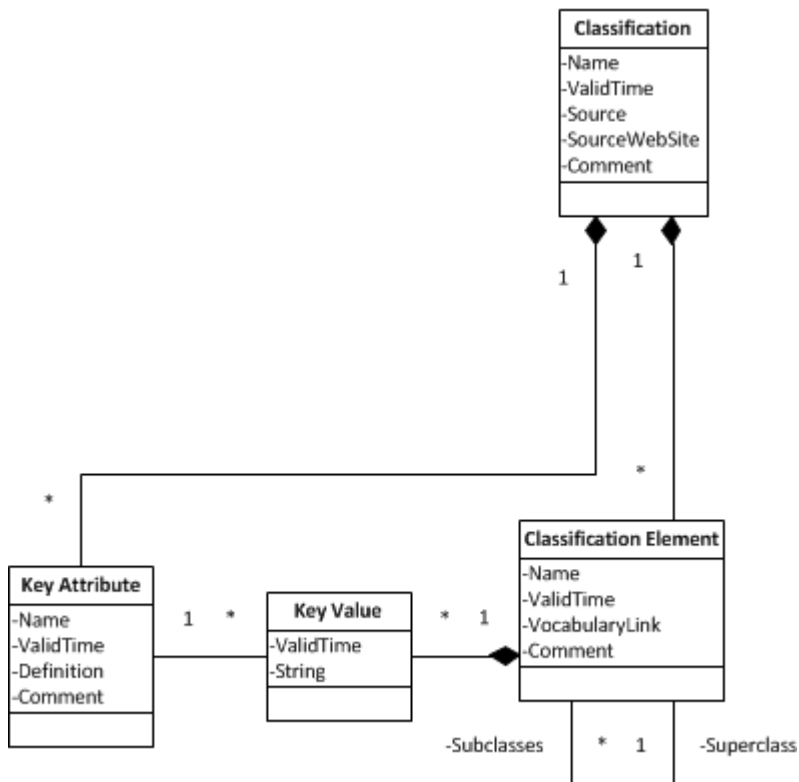
De Classifications view beschrijft de kennis over Classifications die in de HortiCube is vastgelegd. De structuur is weergegeven in Figuur 3.

Een **Classification** bevat een opsomming van Classification Elements die als Dimension in Data Sets gebruikt kunnen worden. Een voorbeeld is een Classification met de Name "Alpha-2 country code". De Source van deze Classification is de "ISO 3166 Standard", beschreven op WebSite [http://www.iso.org/iso/country\\_codes](http://www.iso.org/iso/country_codes).

Een **Classification Element** heeft een Name, bijvoorbeeld "Peren" en kan de URI bevatten van een plaats op het World Wide Web, waar de definitie gevonden kan worden. Een Classification kan een hiërarchie bevatten, bij voorbeeld "Peren" kan verbonden zijn met Superclass "Pitvruchten" en er kan via Subclasses een aantal andere Classification Elements onder vallen, waaronder bijvoorbeeld "Conference", "Doyenné du Comice" en "Williams' Bon Chrétien".

Aan de Classification Elements wordt vanuit de Data Sets meestal gerefereerd met een code, bestaande uit (meestal) één of (soms) een combinatie van **Key Values**. De "Alpha-2 Country Code", bijvoorbeeld, is een enkelvoudige code, bestaat uit twee letters die een land identificeren, bijvoorbeeld "FR" voor Frankrijk. De manier waarop variëties van fruit in de GPC zijn gecodeerd, is een voorbeeld van een samengestelde code, waarin bijvoorbeeld de Brick code (10005903) verwijst naar de soort "Pears", de Attribute code (20002795) verwijst naar het attribute "Pears Variety", en een Attribute Value code (30016335) verwijst naar de variety "Conference".

Een Classification heeft een set **Key Attributes**, die overeenkomen met Dimensions van een Data Set. De set Names van Key Attributes voor identificatie van een variety van peren uit bovenstaand voorbeeld uit de GPC is {"Brick Code", "Attribute Code", "Attribute Value"}. Voor gebruik in de HortiCube en de worden van elk Key Attribute (naast een Name) ValidTime, Definition en eventueel Comment vastgelegd.



Figuur 3. Classifications view



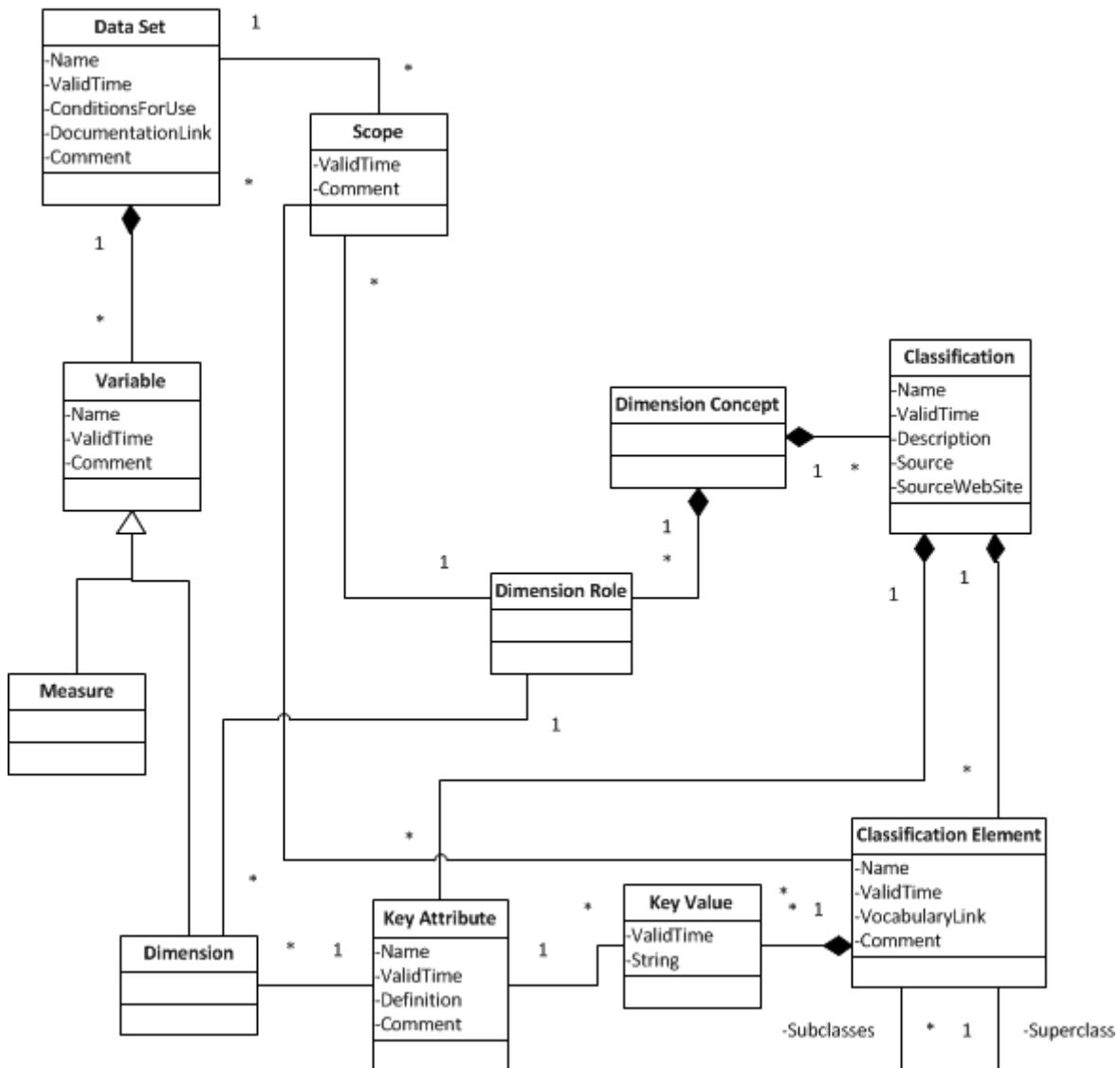
#### 4. Scopes view

De Scopes view beschrijft de inhoud van Data Sets in termen van de Classification Elements waarover data zijn opgenomen. De structuur is weergegeven in Figuur 5.

De **Scope** van een Data Set is de set Classification Elements die kunnen voorkomen in een bepaalde Dimension van de Data Set. Bijvoorbeeld: de importstatistiek van Eurostat (COMEXT) heeft een dimensie "Declarerend Land". Die kan via de rol "Bestemmingsland" en het Concept "Land" verbonden zijn met een landenclassificatie die alle landen van de wereld bevat. De Data Set bevat echter alleen de lidstaten van de EU als "Declarerend Land" voor importen.

Een Data Set kan een beperkte scope hebben voor elk van de Dimension Roles die erin gerepresenteerd zijn.

Een Scope geeft informatie over de betekenis van het ontbreken van data. Als in een Data Set met de EU als scope data over importen van België van een bepaald product ontbreken, kan dat geïnterpreteerd worden als "er heeft in de betreffende periode in België geen import van dit product plaatsgevonden". Als in dezelfde Data Set data van Australië (dat wel voorkomt in de gebruikte landenclassificatie) ontbreken, betekent het niets: Australië valt buiten de Scope.

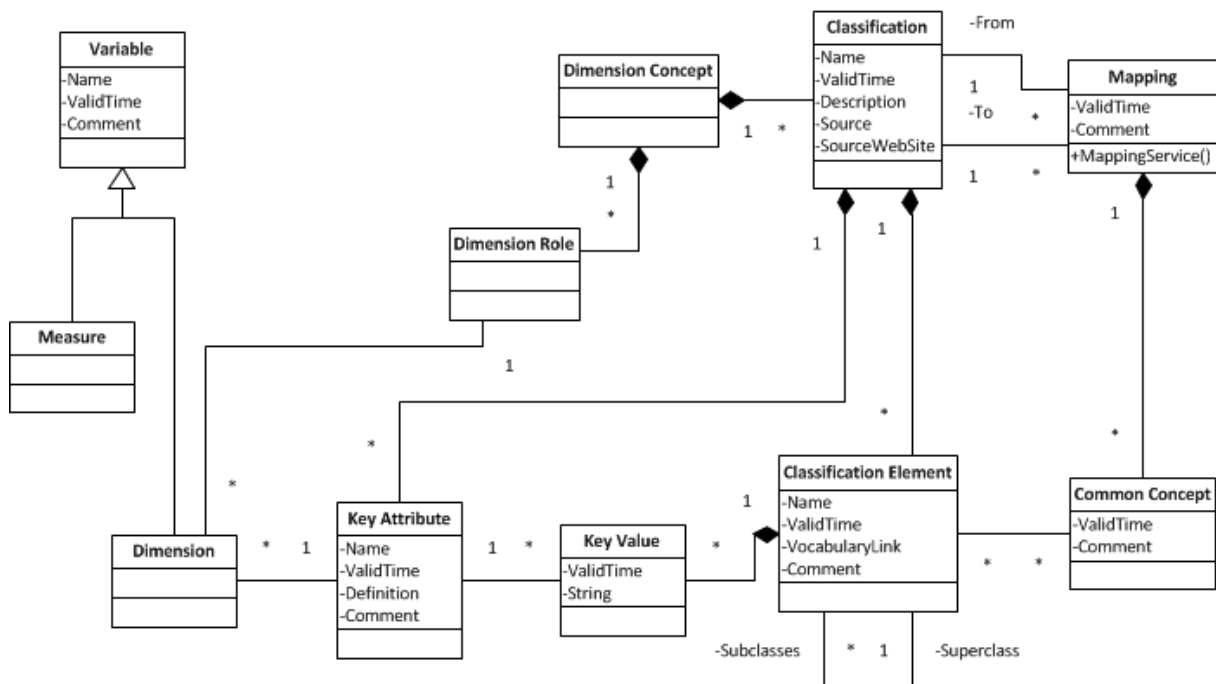


Figuur 5. Scopes view

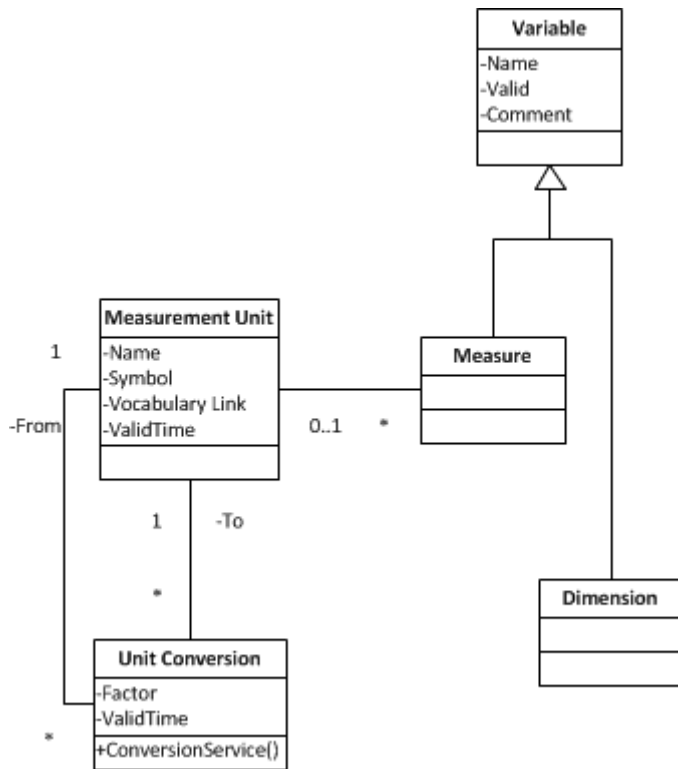
## 5. Conversions en Mappings

Het kennismodel bevat informatie waarmee de HortiCube conversies en mappings kan uitvoeren:

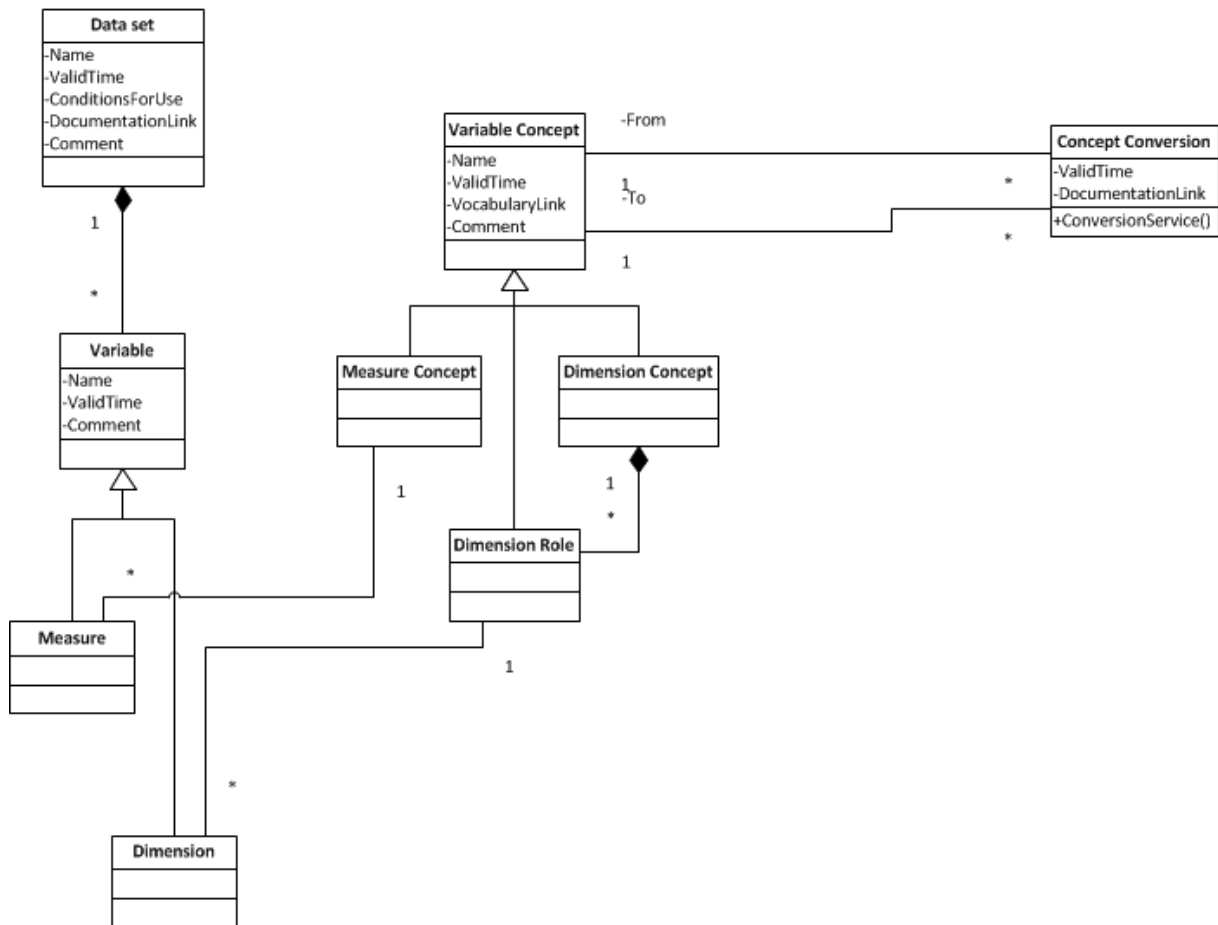
- **Mapping** van Classifications (zie Figuur 6): Een Mapping is verbonden met twee Classifications (From en To) en kent een MappingService die de mapping uit kan voeren. Het Mapping is gebaseerd op het principe van **Common Concepts**. Bij elk Classification Element wordt een VocabularyLink opgenomen, een URI die verwijst naar een definitie van het Classification Element, ergens op het Worldwide Web. Als twee elementen dezelfde VocabularyLink hebben, dan zijn de Elementen te mappen. De set bij een Mapping behorende Common Concepts wordt overigens niet noodzakelijk vastgelegd, maar kan telkens “on the fly” opgebouwd worden als een Mapping uitgevoerd moet worden.
- **Unit Conversion** (zie Figuur 7): omrekening van eenheden. Soms is dat eenvoudig te doen met een vaste Factor, zoals van kilogram naar ton. Soms is een ingewikkelder algoritme nodig, zoals bij valuta. De Unit Conversion kent ook de ConversionService die het werk kan doen.
- **Concept Conversion** (zie Figuur 8): het berekenen van de waarde van een Variabele uit een of meer andere Variables, bijvoorbeeld berekening van gemiddelde prijs uit waarde en hoeveelheid. Ook de Concept Conversion kent een ConversionService die het werk kan doen.



Figuur 6. Mapping



Figuur 7. Unit Conversion



Figuur 8. Concept Conversion

### Bijlage 1: Overzicht van het domeinmodel van metadata in de HortiCube

