SAMENVATTING

Achtergrond

In de omgevingswetenschappen wordt veel gebruikgemaakt van computermodellen om beleidsbeslissingen te ondersteunen en om onze complexe omgeving beter te kunnen begrijpen. Deze modellen zijn als het ware abstracte weergaves van systemen in de natuur. De wetenschappers die deze modellen ontwikkelen beschikken over domeinkennis, dat wil zeggen dat zij goed voor ogen hebben welke processen en factoren een rol spelen in hun onderzoeksdomein en hoe deze onderling zijn gerelateerd. Wanneer zij hun modellen ontwikkelen maken deze wetenschappers keuzes welke aspecten zij mee nemen, hoe zij deze vertalen in wiskundige variabelen en vergelijkingen, en hoe zij deze vervolgens implementeren in software. Op die manier wordt hun domeinkennis als het ware verstopt in de broncode van de computermodellen. Deze domeinkennis is echter essentieel om de resultaten en inzichten te begrijpen die voortkomen uit de model berekeningen. Dit is de reden dat computermodellen vaak alleen gebruikt kunnen worden door de mensen die deze zelf hebben gemaakt. Dit proefschrift richt zich op computermodellen die geimplementeerd zijn als spreadsheets. Spreadsheets worden veel gebruikt in de natuurwetenschappen. De domeinkennis van de makers is impliciet weergegeven in de structuur en inhoud van de spreadsheet tabellen. Het doel van dit onderzoek is uit te vinden in welke mate het mogelijk is de domein kennis in natuurwetenschappelijke spreadsheets expliciet te maken. We modelleren de domein kennis op het niveau van het knowledge level, onafhankelijk van de spreadsheet implementatie en onafhankelijk van de taak, en zullen eraan refereren als het domein model.

Aanpak

In dit proefschrift hebben wij onderzocht hoe het domein model het beste kan worden beschreven en gereconstrueerd. Ons onderzoek heeft een exploratief karakter en is gebaseerd op de analyse van een beperkt aantal case studies. Deze case studies bestaan uit spreadsheets, en andere typen computermodellen, die gebruikt

zijn in bestaande onderzoeksprojecten in de natuurwetenschappen. Als eerste hebben we handmatig geanalyseerd hoe patronen in de tekst, structuur en formules in spreadsheet tabellen inzicht kunnen geven in de semantiek. Onze bevindingen hebben we vastgelegd in heuristieken. Deze heuristieken heben we vervolgens gecombineerd met kennis uit externe, gedeelde vocabulaires, i.e., een vocabulaire over het betreffende wetenschapsdomein, en een vocabulaire over eenheden en grootheden, om de inhoud van de spreadsheet tabellen automatisch te annoteren. Deze annotaties maken duidelijk welke entiteiten, instanties en rollen besloten liggen in een spreadsheet tabel. Daarnaast hebben we alle reken relaties weergegeven in een cell dependency graph, i.e., een netwerk dat weergeeft welke spreadsheet cellen via wiskundige vergelijkingen met elkaar zijn verbonden. Met behulp van automatische analyse van deze vergelijkingen en de eerder geformuleerde heuristieken, hebben we de informatie uit dit netwerk vervolgens kunnen aggregeren. Het resulterende, geaggregeerde netwerk bevat alleen die grootheden die een rol spelen in de berekening van model resultaten, en geeft inzicht in hoe de entiteiten en instanties uit de tabel via rekennrelaties met elkaar zijn verbonden. We hebben de verschillende stappen in het reconstructie proces afzonderlijk geevalueerd door voor elke stap onze automatisch gegenereerde resultaten te vergelijken met een handmatig gecreerd 'antwoord model'. Als laatste hebben we een breder, verkennend onderzoek uitgevoerd naar de vorm en inhoud van bestaande natuurwetenschappelijke spreadsheets en ingeschat in hoeverre onze methode bruikbaar is in de praktijk.

Resultaten

In dit proefschrift tonen wij aan dat een adequate bescrijving van het domein model een combinatie bevat van statische informatie, met daarin de entiteiten, instanties en eigenschappen, en dynamische informatie, met daarin hoe deze entiteiten verbonden zijn door reken relaties. De structuur van spreadsheet tabellen bevat impliciete doch waardevolle informatie over het domein model. Spreadsheet tabellen bestaan uit rechthoekige blokken van cellen die semantisch gerelateerd zijn. Deze blokken worden gekarakteriseerd door zowel hun inhoud en positie, als hun rol in de tabel. We hebben vier typen blokken onderscheiden en deze beschreven met overkoepelende concepten uit de OM ontologie voor eenheden en grootheden: 1) Measure (observatie), 2) Unit of Measure (eenheid),

3) Quantity (grootheid), and 4) Phenomenon (fenomeen; object, gebeurtenis of substantie). De Phenomenon blokken in een tabel komen overeen met de entiteiten uit het domein model, en de Quantity blokken met de eigenschappen van deze entiteiten. De groepering van Phenomenon cellen geeft aan welke instanties tot eenzelfde klasse behoren, wat op die manier de hierarchie van de entiteiten bepaalt. De kruislings geplaatste Phenomenon blokken en Quantity instanties laten zien hoe deze blokken via eigenschaps relaties met elkaar zijn verbonden. De grootheden in een tabel fungeren als de verbindende factoren tussen enerzijds de statische domein informatie en anderzijds de dynamische informatie over de berekeningen. We laten zien welke opeenvolgende stappen nodig zijn om het domein model van een natuurwetenschappelijke spreadsheet te reconstrueren, en voor elke stap geven we aan in hoeverre we deze automatisch kunnen ondersteunen. We hebben ondervonden dat, in de praktijk, het merendeel van de spreadsheets niet voldoet aan onze eisen betreffende hun structuur en inhoud. Door ons automatische proces aan te vullen met heuristieken en kennis uit externe gedeelde vocabulaires zijn we toch in staat het domein model van bestaande spreadsheets te reconstrueren, ook al bevatten de tabellen geen complete informatie. Als laatste resultaat hebben we richtlijnen opgesteld voor het ontwerpen van natuurwetenschappelijke spreadsheets. Deze richtlijnen informeren natuurwetenschappers hoe ze de kennis in hun spreadsheets expliciet kunnen maken, zonder hen al te veel beperkingen op te leggen in het ontwerp proces. De richtlijnen lenen zich ook voor implementatie in software om zo automatisch het ontwikkel proces te kunnen ondersteunen.

Conclusies

In dit proefschrift hebben we aangetoond dat het modelleren van kennis opgesloten in natuurwetenschappelijke computermodellen een geschikte manier vormt om deze kennis expliciet te maken. Het annoteren van de model code en data met concepten uit externe, gedeelde vocabulaires voegt context en betekenis toe. Op deze manier kunnen de model code en data worden begrepen en gebruikt door meer mensen dan alleen de oorspronkelijke makers. Sterker nog, de annotaties veranderen opzichzelf staande model code en data in linked data, die gedeeld en hergebruikt kunnen worden in het Semantisch Web.