Stichting NIOC en de NIOC kennisbank

Stichting NIOC (www.nioc.nl) stelt zich conform zijn statuten tot doel: het realiseren van congressen over informatica onderwijs en voorts al hetgeen met een en ander rechtstreeks of zijdelings verband houdt of daartoe bevorderlijk kan zijn, alles in de ruimste zin des woords.

De stichting NIOC neemt de archivering van de resultaten van de congressen voor zijn rekening. De website www.nioc.nl ontsluit onder "Eerdere congressen" de gearchiveerde websites van eerdere congressen. De vele afzonderlijke congresbijdragen zijn opgenomen in een kennisbank die via dezelfde website onder "NIOC kennisbank" ontsloten wordt.

Op dit moment bevat de NIOC kennisbank alle bijdragen, incl. die van het laatste congres (NIOC2023, gehouden op donderdag 30 maart 2023 jl. en georganiseerd door NHL Stenden Hogeschool). Bij elkaar bijna 1500 bijdragen!

We roepen je op, na het lezen van het document dat door jou is gedownload, de auteur(s) feedback te geven. Dit kan door je te registreren als gebruiker van de NIOC kennisbank. Na registratie krijg je bericht hoe in te loggen op de NIOC kennisbank.

Het eerstvolgende NIOC vindt plaats op donderdag 27 maart 2025 in Zwolle en wordt dan georganiseerd door Hogeschool Windesheim. Kijk op www.nioc2025.nl voor meer informatie.

Wil je op de hoogte blijven van de ontwikkeling rond Stichting NIOC en de NIOC kennisbank, schrijf je dan in op de nieuwsbrief via

www.nioc.nl/nioc-kennisbank/aanmelden nieuwsbrief

Reacties over de NIOC kennisbank en de inhoud daarvan kun je richten aan de beheerder:

R. Smedinga kennisbank@nioc.nl.

Vermeld bij reacties jouw naam en telefoonnummer voor nader contact.

Data Vault & BI

Data Vault in het bacheloronderwijs

Een alternatieve visie op data warehouse modellering en architectuur.

Door: Dineke Romeijn, Hogeschool Arnhem en Nijmegen.

Trefwoorden: Data Vault, data warehousing, Business Intelligence, gegevensmodellering, datawarehouse-architectuur.

In het bedrijfsleven neemt de belangstelling voor Business Intelligence en data warehousing nog steeds toe. Jarenlang domineerden binnen data warehousing grofweg twee benaderingen en mengvormen hiervan. De afgelopen jaren is daar een andere benadering bijgekomen, de Data Vault. Data Vault staat voor zowel een visie op datawarehouse-architectuur als op het gegevensmodel voor dat data warehouse. Deze benadering is in Nederland binnen korte tijd zeer populair geworden. Bij de HAN is deze benadering opgenomen in het curriculum van de minor Business Intelligence. Onze ervaringen hiermee deelden we met het publiek tijdens deze presentatie.

Wat is Data Vault?

Data Vault is een gegevensmodel voor een DataWareHouse (DWH) en een architectuur. Het model is ontworpen door Dan Linstedt in 2002. Volgens Linstedt is het: “A collection of detail-oriented, history-tracing and uniquely linked collection of normalized tables” én “A hybrid approach combining the best of 3NF and dimensional modelling”. In relatie tot bedrijfsvoering noemt Linstedt het: “Flexible, scalable, consistent and adjustable to business needs” en in staat “to handle high-speed data loads in parallel”.

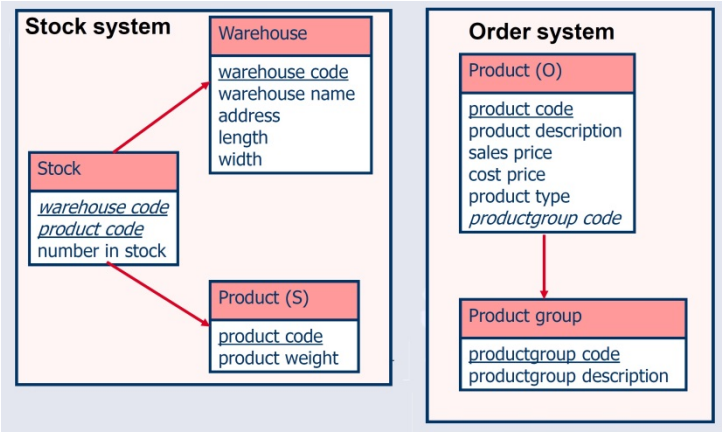
Het Data Vault model onderscheidt drie componenten:

- Hubs: bevatten de bedrijfssleutels van entiteiten;
- Links: relaties tussen entiteiten, afhankelijke entiteiten;
- Satellieten: alle andere relevante attributen van een entiteit of relatie.

Daarnaast bestaan er surrogaatsleutels voor hubs, eventueel ook voor links; is er voor alle elementen een verwijzing naar de bron en een timestamp voor de laaddatum en is voor satellieten de timestamp deel van de sleutel (vergelijkbaar met SCD van Kimball).

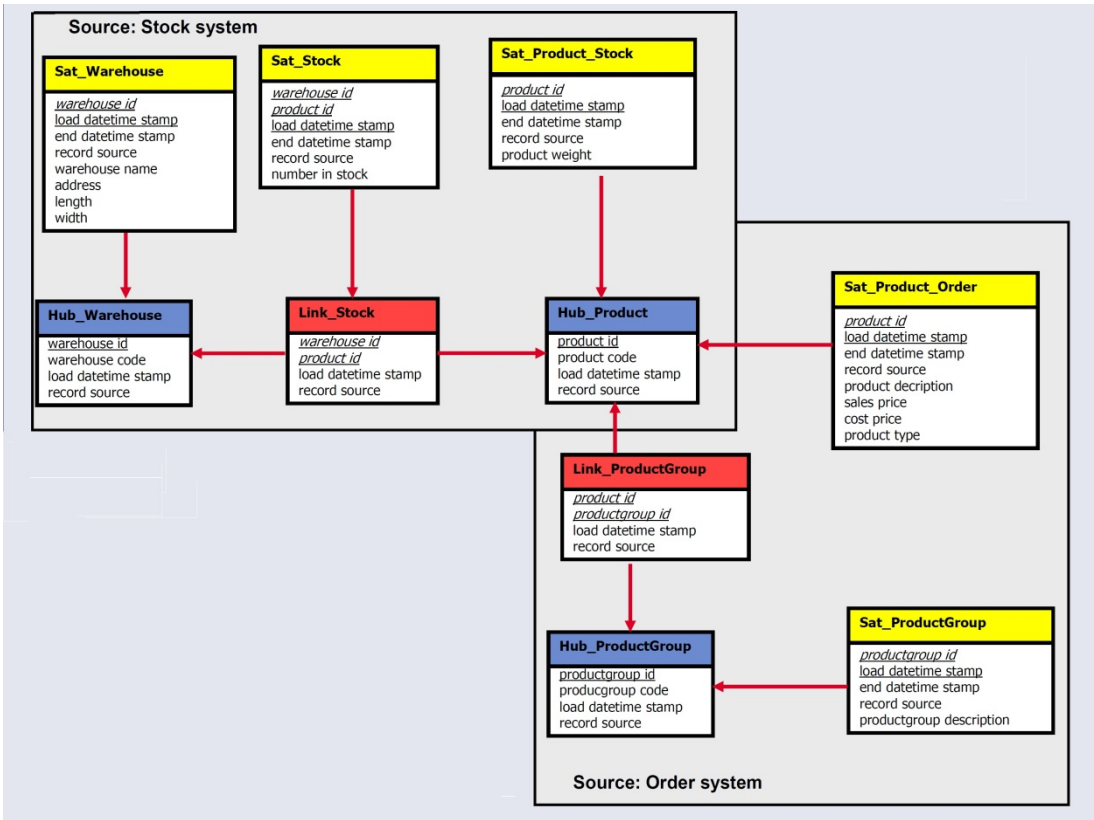
Data Vault: een voorbeeld

Figuur 1 toont een voorbeeld van twee bronsystemen Voorraad (Stock) en Bestellingen(Order). Het Stocksysteem bestaat uit drie relationele tabellen (Warehouse, Stock en Product) met twee relaties (via warehousecode en productcode) en het Ordersysteem bevat twee relationele tabellen (Product en Productgroup) met een relatie via productgroupcode.

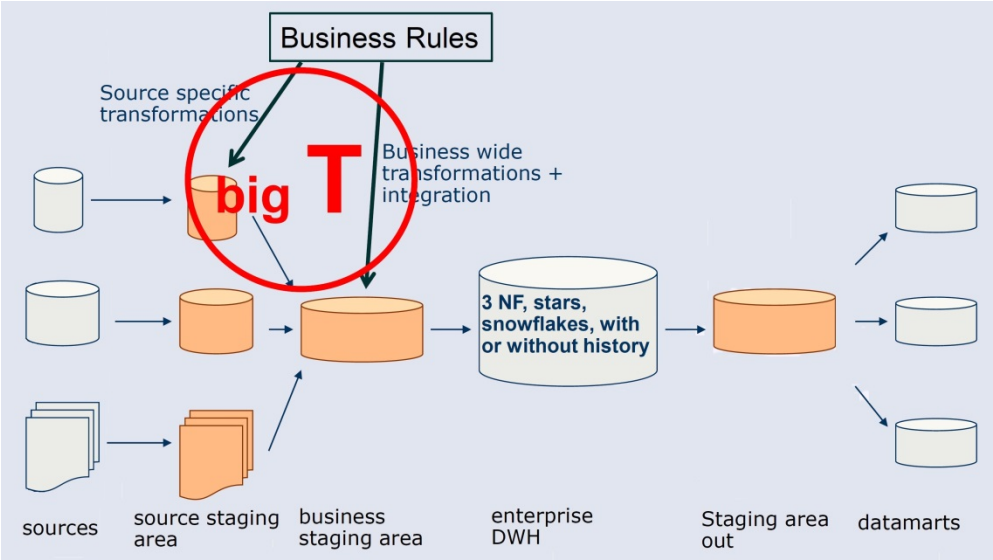


Figuur 1. Data Vault voorbeeld: twee bronsystemen (Stock en Order).

In het Datavault systeem wordt van elke brontabel een Satelliettabel én een Hubtabel of Linktabel afgeleid. De satelliettabel bevat de velden van de brontabel met toegevoegde velden: load_datetime_stamp, end_datetime_stamp en record_source. De primaire sleutel is vervangen door een bron-record_id. Bijvoorbeeld de warehouse-code is vervangen door warehouse-id. De Hubtabel Hub_Warehouse bevat de velden: warehouse-id, warehouse-code, load_datetime_stamp en record_source. De linktabel Link_Stock bevat in dit voorbeeld de identificatievelden warehouse_id en product_id en daarnaast de velden load_datetime_stamp en record_source van de bijbehorende satelliet (Sat_Stock). Voor het bronsysteem order zijn twee Hubtabellen en een Linktabel afgeleid. In dit voorbeeld worden de beide bronsystemen verbonden door de 'gezamenlijke' Hubtabel Hub_Product (figuur2).

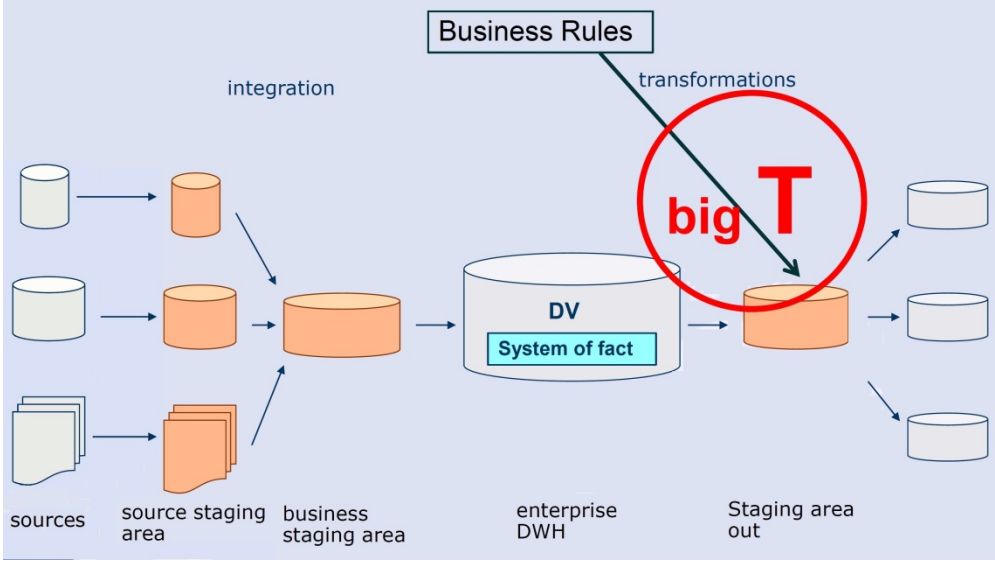


Figuur 2. DataVault voorbeeld: afgeleide Hub- en Link-tabellen uit bronsystemen (Stock en Order).



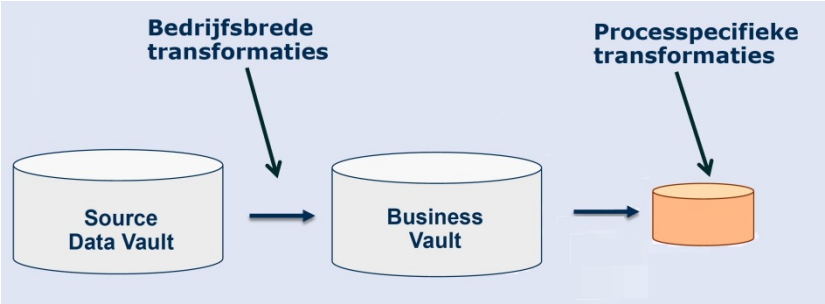
Figuur 3. Klassieke Architectuur voor Datawarehousing.

De klassieke architectuur voor Datawarehousing benut een staging area voor bron-specifieke transformaties op basis van bedrijfsregels en bundelt de bronnen via een tweede transformatie samen in een business staging area (figuur 3). Daaruit worden DWH-modellen (3NF, stars of snowflakes) met of zonder historie afgeleid. De uitvoer staging area produceert datamarts voor de gewenste rapportages.



Figuur 4. Data Vault Architectuur.

De Data Vault Architectuur integreert zonder transformaties de source staging area en de business staging area in een enterprise datawarehouse als feitensysteem. De bedrijfsregels bepalen nu de transformaties voor de realisatie van de uitvoer staging area, waaruit de datamarts voor gewenste rapportages worden afgeleid (figuur 4).
Figuur 5 toont een uitgebreide Data Vault architectuur. Daarbij wordt een Data Vault uit de bronsystemen door bedrijfsbrede transformaties omgevormd tot een Bedrijf Data Vault. Daaruit kunnen door proces specifieke transformaties datamarts worden afgeleid die proces specifieke rapportages kunnen leveren.



Figuur 5. Uitgebreide Data Vault Architectuur

Voordelen en nadelen in de beroepspraktijk

De voordelen van de Data Vault Architectuur in de beroepspraktijk zijn divers. De opslag van de complete historie is mogelijk en dat versterkt en verruimt de mogelijkheden voor compliance en auditibility. Het model kent 3 simpele elementen en is flexibel en schaalbaar.

Data Vault bevat uitsluitend feiten (system of fact) zoals bekend in de bronnen en is daardoor volledig bronafhankelijk en *niet* regelafhankelijk. Daarmee kunnen verantwoordelijkheden van IT en business goed gescheiden worden. Model-gedreven generatie van DataVault systemen is mogelijk met 'parallel laden' en snelle genereerbare ETL. Als gevolg zijn de ontwikkel- en onderhoudskosten lager. Er is een ware 'explosie van tooling' op het gebied van Data Vault. Het is een belangrijk onderwerp van discussie in BI Dutch' (de Nederlandse Business Intelligence Community op LinkedIn) en er zijn sinds 2011 meer dan 15 tools beschikbaar.

De nadelen van de DataVault Architectuur hebben te maken met een gebrek aan standaardisatie door het ontbreken van een heldere formele beschrijving. Lindstedt kan als 'gesloten bedenker' in het dynamische en brede belangstellingsveld niet alle discussiepunten (op)volgen en dat leidt tot variaties en discussie over best practices. Dat kan technisch leiden tot extra (evt. virtuele) lagen en extra modellen. ETL voor DataVault is complexer bij handmatige ontwikkeling en daarom zijn generatoren essentieel, maar de ontwikkeling daarvan is in de geschetste praktijk niet eenvoudig. Geen nadeel maar ook geen voordeel is dat aan de 'big T' (de standaard specificatie van de complexe transformatie op basis van bedrijfsregels) niemand de vingers wil branden, ook Data Vault niet….

Business Intelligence in het curriculum van de HAN

Opname van Data Vault in curriculum is van belang door de toenemende vraag naar analyse van gegevens en is duidelijk sprake van een 'data-explosie'. De schatting van sommige consultants is, dat 80 % van DWH projecten (voor nieuwe of herziening van bestaande oplossingen) gebruik maakt van Data Vault. Volgens anderen is dat wat optimistisch, maar er lijkt toch een toenemend gebruik waarneembaar in de beroepspraktijk. Hoewel er andersoortige oplossingen bij komen, blijft klassiek datawarehousing in de praktijk nog steeds zeer belangrijk. Figuur 6 toont de ontwikkeling van BI-onderwijs in het curriculum van de HAN opleidingen.

Periode	Business Intelligence in ICT – curricula van de HAN
1998-2002	Keuzemodule Datawarehousing; Business Intelligence voor de Hogere Informatica opleiding.
2002-2008	Business Intelligence in het curriculum voor BIM, I en TI opleidingen.
2003-2012	Business Intelligence in de masteropleiding 'Information Systems Development'.
2009-	Minor Business Intelligence.

Figuur 6. Ontwikkeling Business Intelligence onderwijs in de HAN-opleidingen.

In de huidige curricula van de ICT-opleidingen van de Informatica Communicatie Academie is Business Intelligence opgenomen in drie semesters: voorbereidend in 'Develop an Information System' (DIS) en 'Align Business and IT' (ABI) en inleidend in het semester Develop an e-Community (DECO). Het zwaartepunt van BI-onderwijs ligt in de minor Business Intelligence en in afstudeeropdrachten bij bedrijven of lectoraten.

De minor Business Intelligence

De minor BI is een onderzoek minor als verdiepende minor voor de opleidingen Business IT en Management (BIM) en Informatica (I). De voorkennis van studenten is wisselend, omdat ze in jaar 2, 3 of 4 kunnen instromen. Het is een verbredende minor voor studenten van de Faculteit Economie en Management (FEM) en de opleiding Technische bedrijfskunde en dat is alleen mogelijk na een voorbereidingstraject. De minor bestaat uit twee courses en een onderzoeksproject (figuur 7).

EC	Onderdeel	Inhoud
7,5	Course 1: Voortraject BI	**Basiskennis**: de Business Intelligence (BI) processen; Doelstellingen voor bedrijfsprocessen; Strategische doelstellingen; (Meet-)gegevens, een meetplan en een actieplan opstellen voor een doelstelling; Presenteren en verspreiden van informatie; BI-projecten, beheer en exploitatie, outsourcing; Business Intelligence implementeren; BI promoten
		Opdracht 1: Een BI advies voor bedrijfsprocessen (doelstellingen, meetgegevens, meetplan, actieplan, business case en implementatieplan opstellen).
		Opdracht 2: Een strategisch BI advies (doelstellingen, strategisch informatieplan opstellen).
	Course 1: Prototypetraject	**Centrale beroepstaken**: Hoe ontwerp je een systeem dat snel, accuraat en flexibel de gewenste informatie levert? Hoe realiseer je vervolgens een prototype van dit systeem?
		Kennis: transactional processing versus analytical processing; datawarehouses, datamarts; BI architecturen; **Data Vault;** sterren en dimensies; historie van gegevens; het ontwikkelproces en ondersteunende methoden; ETL;
		Casus: Integreert voortraject en prototypetraject; Ontwerp en realiseer een BI omgeving
		Leermiddelen: Literatuur: "Sterren en dimensies", H. van der Lek, F. Habers, M. Schmitz; "Business Intelligence en datawarehousing", 2e editie, K. Verhagen; Diverse artikelen (op HAN-ELO). Software: Microsoft Excel 2007/2010; SQL Server 2008 R2, eventueel 2005 (incl. service pack 3); SQL Server Business Intelligence Studio; Power Designer; Front-end tool naar keuze.
7,5	Course 2: BI-thema's	Diverse BI thema's; Literatuuronderzoek; Capita selecta; Gastcolleges;
15	Project BI	Een in overleg gekozen onderzoek uitvoeren; Start met onderzoek vaardigheden.

Figuur 7. Onderdelen en inhoud van de minor Business Intelligence.

Ervaringen met Data Vault in BI minor

In 2008 is een eenmalige op zichzelf staande les over het onderwerp Data Vault in de course opgenomen. Aanvullend is een losstaande workshop van een halve dag en een aanvullende opdracht bij de casus toegevoegd en momenteel is Data Vault volledig geïntegreerd in casus en lessen en de 'centrale methode' voor modelleren van Datawarehouses.

In de BI minor worden nu drie aspecten van Data Vault behandeld. Als eerste het Data Vault model als gegevensmodel voor een DWH. Daarvoor worden twee dagdelen gebruikt aan het begin van prototypetraject met inleidende presentaties en gerichte oefeningen met kleine voorbeelden. Via vaardigheidstraining moeten studenten een goede basis leggen om Data Vault als integraal

onderdeel van de casus toe te passen. Daarin zijn complexere modelleerkwesties verwerkt. Met een goede basis ontdekken de meeste studenten dit zelf. Studenten krijgen nadere uitleg aan de hand van hun vragen en door bespreken van alternatieven.
Als tweede aspect de Data Vault architectuur die vanaf week 1 steeds een terugkerend thema is en daarbij oplopend in complexiteit wordt behandeld. Het begrip neemt toe, maar landt pas echt goed na diepgaand inzicht in BI, meestal aan eind van de minor. Dit onderwerp wordt deels vraag gestuurd behandeld, aansluitend op de lesstof komen veel studenten met vragen om nadere uitleg, dan worden ook alternatieven besproken. Inzicht verwerven in architectuur blijkt veel lastiger dan inzicht in het Data Vault model.
Als derde aspect komt model gedreven generatie aan de orde in een gastpresentatie en workshop: 'Genereren van sterschema's en Data Vault via tooling'. Een praktijkvoorbeeld uit bedrijfsleven (QOSQO) wordt gepresenteerd door een Ms ISD afgestudeerde (Lulzim Bilali) die als onderzoeker in het lectoraat MBIS heeft gewerkt. Daarbij wordt een open source tool (Quipu) toegepast. Afsluitend wordt een presentatie vanuit het lectoraat over model gedreven generatie gegeven: 'transformatie vanuit een conceptueel model naar Data Vault'. Een knelpunt bij dit aspect dat extra aandacht vereist, is de sterk verschillende voorkennis van de studenten.
Voor de BIM studenten met geringe ervaring en oefening in scripting of programmeren (in SQL) is ETL lastig. Door goede voorbeelden, gebruik van tooling en het nut en belang benadrukken van altijd bruikbare degelijke SQL-kennis en vaardigheid kan de inzet en motivatie worden behouden.
De voorkennis van studenten is verschillend, ze zijn niet allemaal bekend met ERM of FOM. Voor de casus is van belang dat deze als één van de eerste onderdelen kan blijven hangen en in belang wordt overschat. Knelpunt in de course uitvoering blijkt ook het ontbreken van één standaard voor Data Vault. De alternatieven maken dan de courseinhoud complexer. Dat is leuk voor goede studenten of studenten met de juiste voorkennis, maar het kan erg verwarrend zijn voor anderen met minder voorkennis en vaardigheid.

Bedrijfscontacten
De bedrijfscontacten rondom de BI-minor en het BI-onderzoek zijn belangrijk om de ontwikkelingen in de beroepspraktijk op de voet te kunnen volgen. Er is een goede relatie en samenwerking met de NLDVGG (Nederlandse Data Vault Gebruikersgroep; Martijn Evers). Ook draagt de organisatie van studiedagen informatiemodellering voor deskundigen in bedrijfsleven en medewerkers van M-BIS bij aan de verbinding met BI toepassende bedrijven. Bij de uitvoering van de courses versterken bedrijfsbezoeken van studenten en gastpresentaties voor studenten ook de relatie met bedrijven. Samen hebben deze relaties geleid tot organisatie van cursussen met en voor bedrijven (MATTER) en organisatieparticipatie in congressen zoals Data Modeling Zone Europe 2013.

Wilt u reageren op deze presentatie? Neem dan contact op met:
Dineke Romeijn-de Jager; docent-onderzoeker; Informatica Communicatie Academie; Lectoraat Model-based Information Systems; Hogeschool van Arnhem en Nijmegen
dineke.romeijn@han.nl