Entiteitreconciliatie

ondanks beperkte overlap

door middel van

objectgelijkheid

Casus

"Koppelen van persoonsgegevens zonder een gemeenschappelijke identificatie"

Justitie Wetenschappelijk Onderzoek- en Documentatiecentrum

Universiteit Twente

de ondernemende universiteit

Scriptie

J.J. van Dijk

Den Haag, augustus 2006

Afstudeercommissie	Dr. Ir. H.E. Blok Dr. M.M. Fokkinga Drs. R.F. Meijer (WODC)
Groep	Databases
Opleiding	Informatica Universiteit Twente, Enschede

Abstract

Het koppelen van informatiebronnen wordt in de huidige maatschappij steeds belangrijker. Door koppelen ontstaan nieuwe inzichten, omdat nieuwe gegevens van gemeenschappelijke objecten met elkaar in verband kunnen worden gebracht. Dit onderzoek richt zich op het koppelen van bronnen op microniveau. Hierbij worden entiteiten, die naar hetzelfde object verwijzen, aan elkaar gekoppeld: *entiteitreconciliatie* (bijvoorbeeld persoonsentiteiten die naar één persoon verwijzen). Verschillende bronnen hebben vaak geen gemeenschappelijke identificatie, waardoor deze manier van koppelen afvalt. Bronnen die interessant zijn om te koppelen, bevatten vaak weinig gemeenschappelijke informatie. Vanwege de beperkte overlap is de winst van het koppelen het grootst; er kunnen meer nieuwe gegevens met elkaar in verband worden gebracht. Overlap is echter, zonder gemeenschappelijke identificatie, wel de enige troef in de poging om te koppelen.

Om ondanks beperkte overlap toch entiteiten te kunnen reconciliëren, is een theorie ontwikkeld om alle aanwezige overlap tussen twee informatiebronnen te gebruiken. Overlap bestaat uit eigenschappen die beide bronnen gemeen hebben. Als een gemeenschappelijke eigenschap overeenkomt, dan is er sprake van gelijkheid. De mate waarin zo'n eigenschap overeenkomt, wordt bepaald door de afstand tussen twee attributen die de eigenschap beschrijven. Met behulp van expertkennis wordt deze afstand via een afstandsverdeling (een trendlijn over het histogram van de verwachte afstanden van de eigenschap) omgezet in een mate van gelijkheid. De attributen, die een gemeenschappelijke eigenschap beschrijven, worden geplaatst onder een gemeenschappelijk entiteittype (knoop genoemd). Elke knoop draagt bij aan de beschrijving van de centrumknoop waarin de reconciliatie gewenst is. Zodoende wordt de entiteitgelijkheid per knoop bepaald en wordt ook de objectgelijkheid bepaald, waarin tevens de gelijkheid van andere knopen wordt meegenomen. Hierbij wordt de gelijkheid effectief gedistribueerd naar de centrumknoop. Door de knopen te berekenen in een hiërarchische structuur ontstaat clustering, waardoor het aantal vergelijkingen wordt verlaagd. Voor de entiteitreconciliatie is een methode bedacht, waarmee entiteiten van één knoop efficiënt worden gereconcilieerd.

Om de theorie te toetsen is een prototype (EROS, '*Entity Reconciliation using Object Similarity*') ontwikkeld, waarin een casus is geïmplementeerd. Van deze casus zijn de correcte reconciliaties bekend; deze zijn gebruikt in de analyse van de resultaten. Er is persoonsreconciliatie toegepast op 10.000 personen in de ene bron tegen 8.705 personen in de andere bron. Hierbij zijn 5 gemeenschappelijke eigenschappen gebruikt, waaronder 3 persoonseigenschappen (geboorteland, geslacht en geboortedatum). Voor 5% is de correcte reconciliatie niet gevonden als gevolg van te weinig overlap. Als de correcte reconciliatie is gevonden, dan wordt deze in 98% van de gevallen ook daadwerkelijk gekozen.

Uit dit onderzoek blijkt dat, ondanks beperkte overlap, reconciliatie op microniveau door middel van objectgelijkheid goed mogelijk is. Uiteraard moet de aanwezige overlap discriminerend genoeg zijn. In dit kader moet worden opgemerkt dat door de kleine set van gegevens de persoonseigenschappen voor sommige combinaties al sterk discriminerend zijn. Meer onderzoek is nodig om te bepalen wanneer overlap voldoende discriminerend is, met name voor grotere datasets waarin de correcte reconciliaties onbekend zijn. De theorie biedt een uitgangspunt voor meer onderzoek in de richting van data mining en privacy-gerelateerde toepassingen.

Voorwoord

Deze scriptie markeert het einde van mijn studie Technische Informatica aan de Universiteit Twente. Voor mij was het Wetenschappelijk Onderzoek- en Documentatiecentrum (WODC) een uitdagende en leerzame plek om het onderzoek voor mijn scriptie uit te voeren. Ik heb veel geleerd over het onderzoeksgebied van mijn scriptie; het koppelen van informatiebronnen. Door de praktijksituatie moest het koppelen gebeuren met zeer weinig en soms lastig te definiëren overlap. Ik heb mij geconcentreerd op het koppelen van één gemeenschappelijk entiteittype, waarbij het mogelijk moet zijn om alle aanwezige overlap tussen de informatiebronnen effectief te gebruiken. Daarnaast heb ik gewerkt aan een manier om de gelijkheid in de overlap efficiënt te berekenen.

Deze scriptie had niet tot stand kunnen komen zonder de hulp van anderen, die ik in dit voorwoord wil bedanken.

Allereerst wil ik het WODC bedanken voor de mogelijkheid om mijn onderzoek uit te voeren. Het WODC is een plek waar wetenschappelijk zeer veel uitdagend werk te doen is. Ik heb de vrijheid gekregen om in de vele mogelijkheden mijn eigen weg te zoeken voor mijn onderzoek. Graag wil ik Dr. Ir. Sunil Choenni bedanken voor de hulp bij het zoeken van die weg. Veel dank ben ik verschuldigd aan Drs. Ronald Meijer – mijn begeleider bij het WODC – voor de grote hoeveelheid tijd die hij in mijn begeleiding gestoken heeft. Zijn interesse en enthousiasme vormden een grote inspiratie voor mij. Daarnaast wil ik mijn begeleiders van de universiteit, Dr. Ir. Henk Ernst Blok en Dr. Maarten Fokkinga, bedanken voor hun waardevolle commentaar.

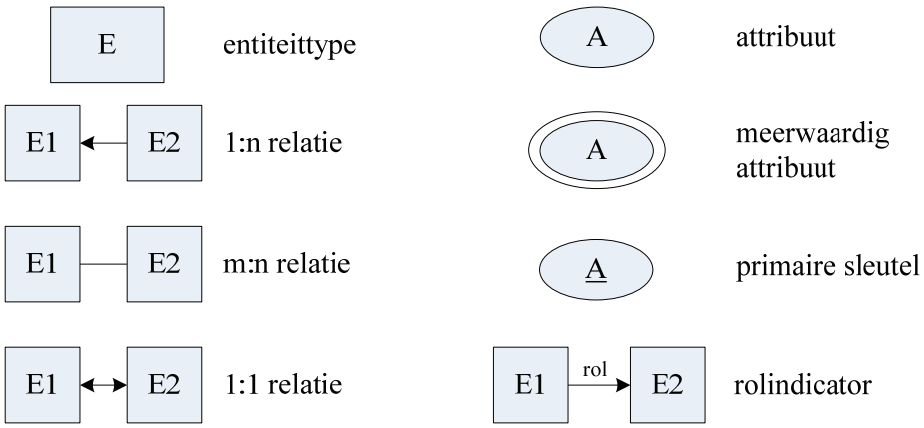
Verder ben ik dank verschuldigd aan de mensen die mij van commentaar hebben voorzien vanuit hun eigen vakgebied of uit interesse; Dolf Trieschnigg (informatica), Nico Alink (wiskunde) en mijn familie hebben zich in een vroeg stadium door de scriptie gewerkt en waardevol commentaar geleverd.

Den Haag, augustus 2006

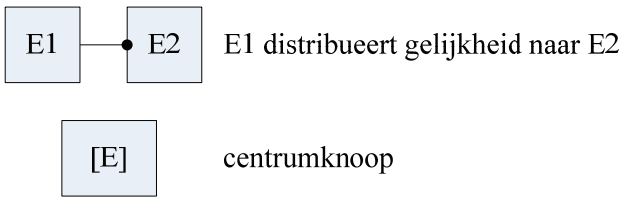
Definities

Begrip	Omschrijving
Attribuut	Een eigenschap of kenmerk van een object, bijvoorbeeld Geboortedatum. De invulling van een attribuut heet de attribuutwaarde.
Centrumknoop	De knoop waarvoor conciliatie gewenst is (zie pagina 15).
Conciliëren / Conciliatie	Het verenigen van de set van entiteiten in een knoop, waarbij objectgelijke entiteiten gereconcilieerd worden (zie pagina 2).
Entiteit	Het voorkomen van een object in een informatiebron in de vorm van een set van attribuutwaarden die hetzelfde object beschrijven. Entiteiten met dezelfde attributen worden gelijksoortige entiteiten genoemd.
Entiteittype	De typering voor de verzameling van gelijksoortige entiteiten in een informatiebron, bijvoorbeeld Persoon.
Knoop	Een gemeenschappelijk entiteittype in twee informatiebronnen (zie pagina 15).
Objectgelijk	Entiteiten of attributen zijn objectgelijk als ze naar hetzelfde object verwijzen.
Reconciliëren / Reconciliatie	Het weer verenigen (koppelen) van objectgelijke entiteiten uit verschillende informatiebronnen (zie pagina 2).
Referentieset	Een representatieve subset van één of meer informatiebronnen.

Notaties in datamodellen



Notaties in gelijkheidsdistributie



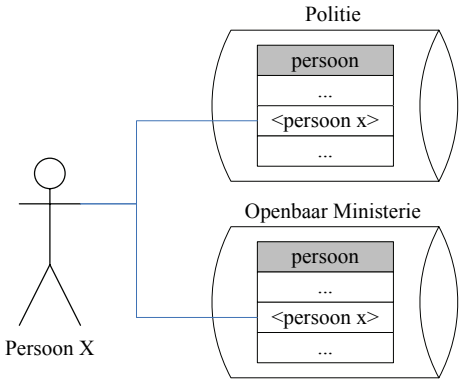
Inhoudsopgave

1 Introductie

1.1 Achtergrond

De wereld bestaat uit objecten zoals personen, gebouwen, etc. Deze objecten kunnen relaties hebben met elkaar. Een persoon bezit bijvoorbeeld één of meerdere gebouwen; gebouwen kunnen personen bevatten. Deze objecten en hun relaties kunnen worden opgeslagen in informatiebronnen. In deze bronnen wordt elk object een (object)entiteit: een voorkomen van het object in een informatiebron.

Een voorbeeld: een persoon X wordt opgepakt voor criminele activiteiten. Er wordt een proces-verbaal (pv) opgemaakt. Hiermee wordt deze persoon in de informatiebron van de politie geregistreerd. Vervolgens wordt het pv tegen persoon X overgedragen aan het Openbaar Ministerie (OM). De gegevens over persoon X en de activiteiten waar de persoon van verdacht wordt, worden ingevoerd in de informatiebron van het OM. Er is geen gemeenschappelijke identificatie, maar persoon X komt in twee informatiebronnen voor:



Figuur 1.1 – Verschillende persoonsentiteiten van één persoon

Zowel de politie als het OM slaan gedetailleerde gegevens van de verdachte, persoon X, op. Dit is nodig voor het uitvoeren van hun taak. Stel nu dat er behoefte is aan onderzoek waarbij de informatie uit beide bronnen op persoonsniveau nodig is. Het is in zo'n geval wenselijk om de persoonsentiteiten, die naar dezelfde persoon verwijzen, te koppelen. In dit geval wordt hiermee de 'loopbaan' van persoon X door de strafrechtsketen in kaart gebracht.

Entiteiten die naar hetzelfde object verwijzen, worden objectgelijke entiteiten genoemd. De methoden om objectgelijke entiteiten te koppelen kunnen worden ingedeeld op twee manieren. De eerste manier is het vinden van voldoende gemeenschappelijke eigenschappen om de entiteiten uniek met elkaar in verband te brengen. Deze gemeenschappelijke eigenschappen vormen dan een gemeenschappelijke identificatie (*sterke sleutel*). Op basis van deze sterke sleutel kunnen objectgelijke entiteiten gekoppeld worden. Helaas is een gemeenschappelijke sterke sleutel niet altijd voor handen.

De tweede manier maakt koppelingen op basis van gemeenschappelijke eigenschappen, zonder dat er sprake is van een sterke sleutel. Voor deze manier bestaan vele benamingen; in dit

document wordt de benaming *reconciliëren* gebruikt, wat 'opnieuw verenigen' betekent[1]. De benaming *reconciliëren* stamt uit de financiële wereld en is door Dey et al ([DSD02]) geïntroduceerd in een wetenschappelijke context.

Voorbeeld. Binnen de debiteurenadministratie worden ontvangen bedragen (de ontvangsten) gereconcilieerd met openstaande posten (facturen). Daarbij wordt een ontvangen bedrag gekoppeld aan één of meerdere openstaande facturen of delen daarvan.

In het voorbeeld worden de ontvangsten gereconcilieerd met de eerder verzonden facturen. Hierbij kan worden aangenomen dat een ontvangst bij een factuur(deel) hoort. Alle ontvangsten dienen te worden gereconcilieerd met een factuur(deel). In andere gevallen is dit niet zo vanzelfsprekend: indien er tegen persoon X onvoldoende bewijs is, dan zal persoon X niet vervolgd worden en daarom niet in de informatiebron van het Openbaar Ministerie terechtkomen. In zo'n geval worden gesproken van de *conciliatie* ('vereniging') van een set persoonsentiteiten, waarbij alleen objectgelijke persoonsentiteiten gereconcilieerd worden.

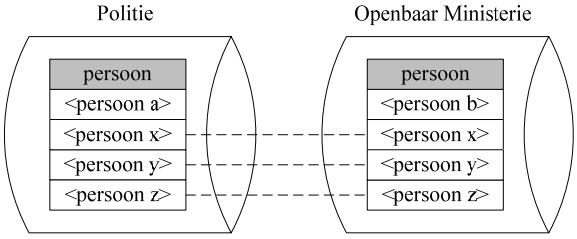
Gegeven twee sets van (persoons)entiteiten P_1 en P_2. Stel dat de functie *obj*(*p*) het object behorende bij een entiteit *p* oplevert.

Definitie. De conciliatie van twee sets van entiteiten P_1 en P_2 is de vereniging $P_1 \cup P_2$ waarvoor geldt:

$$\forall p_1 \in P_1 : \quad p_1 \in P_1 \cap P_2 \quad \leftrightarrow \quad \exists! p_2 \in P_1 \cap P_2 : obj(p_1) = obj(p_2)$$
$$\forall p_2 \in P_2 : \quad p_2 \in P_1 \cap P_2 \quad \leftrightarrow \quad \exists! p_1 \in P_1 \cap P_2 : obj(p_1) = obj(p_2)$$

Definitie. De reconciliatie van twee entiteiten p_1 en p_2 is de vaststelling dat $obj(p_1)=obj(p_2)$.

Voorbeeld. Figuur 1.2 toont de conciliatie van twee sets van persoonsentiteiten in twee verschillende bronnen, waarbij de objectgelijke persoonsentiteiten gereconcilieerd zijn. Persoon A en persoon B hebben geen objectgelijke entiteit in beide sets en worden daarom niet gereconcilieerd.



Figuur 1.2 – Conciliatie van twee sets van persoonsentiteiten

Binnen een informatiebron vallen entiteiten onder een entiteittype. Zo vallen persoonsentiteiten onder het entiteittype Persoon. Twee entiteiten van hetzelfde entiteittype worden *gelijksoortig* genoemd. Elk entiteittype heeft een aantal attributen; geboortedatum en geslacht zijn bijvoorbeeld persoonsattributen. De mate waarin persoonsattributen van twee entiteiten overeen komen, wordt *entiteitgelijkheid* genoemd.

De conciliatie in Figuur 1.2 heeft betrekking op het object Persoon. Een persoon komt als entiteit voor in beide bronnen, maar de entiteiten hebben verschillende eigenschappen. Toch bestaan er vaak verbanden tussen informatiebronnen. Zo hebben personen in beide bronnen een

[1] Reconciliëren: (1) verzoenen, weer verenigen (Van Dale Lexicografie bv, www.vandale.nl, 2006)

delictverleden en komt een delict van een persoon pas in de “Openbaar Ministerie”-bron *nadat* het delict in de “Politie”-bron terecht is gekomen. Al deze eigenschappen kunnen gebruikt worden om te bepalen of twee persoonsentiteiten naar dezelfde persoon verwijzen. Wanneer ook attributen van andere entiteittypen worden vergeleken, dan wordt dit *objectgelijkheid* genoemd.

1.2 Aanleiding

Het Wetenschappelijk Onderzoek- en Documentatiecentrum (WODC) van het Ministerie van Justitie is een onderzoekcentrum dat onderzoek doet met behulp van justitiële informatie-bronnen. De strafrechtsketen – van verdenking tot vervolging en berechting – speelt hierbij een centrale rol. Er is dan ook momenteel veel aandacht voor informatie over daders door de strafrechtsketen heen. Een goed voorbeeld hiervan is de aandacht voor veelplegers. De mogelijkheid om veelplegers te kunnen volgen door de strafrechtsketen levert veel informatie over deze personen. Er is daarom vraag naar persoonsreconciliatie tussen de belangrijkste bronnen in de strafrechtsketen. Deze bronnen zijn externe onderzoeksbronnen, met name geschikt voor onderzoek naar delicten; er bevindt zich daardoor weinig persoonsinformatie in de bronnen.

1.3 Probleemstelling

Als twee informatiebronnen objectgelijke entiteiten bevatten, dan is het waardevol om deze te reconciliëren. Hiervoor is gemeenschappelijke informatie (‘overlap’) nodig. Hierbij is het mogelijk dat de overlap in het bijbehorende entiteittype te beperkt is. De overlap bij andere gerelateerde entiteittypen kan dan worden ingezet.

De centrale probleemstelling luidt nu:

“Hoe kunnen objectgelijke entiteiten gereconcilieerd worden ondanks beperkte overlap?”

Deze probleemstelling kan worden onderverdeeld in een aantal onderzoeksvragen. Bij reconciliëren van objectgelijke entiteiten wordt gebruik gemaakt van de overlap tussen de informatiebronnen. Hieruit komen twee onderzoeksvragen voort:

- Hoe kan de overlap tussen twee informatiebronnen gedefinieerd worden?
- Hoe kan de overlap gebruikt worden in het reconciliëren van objectgelijke entiteiten?

Ondanks een beperkte overlap wordt gezocht naar mogelijkheden om een gemeenschappelijk entiteittype te conciliëren. Hierbij rijst de vraag of er voldoende overlap is om een kwalitatief goede conciliatie te maken. Om dit te kunnen beoordelen is zowel de kwaliteit van de gebruikte gegevens als de kwaliteit van de uiteindelijke conciliatie van belang;

- Hoe kan de kwaliteit van de conciliatie en de gebruikte gegevens uit informatiebronnen bepaald worden?

De kwaliteit van de gebruikte gegevens is afhankelijk van de hoeveelheid inconsistente en missende gegevens. Wat onder de kwaliteit van de conciliatie wordt verstaan, wordt in dit onderzoek verder uitgewerkt. Het verbeteren van de kwaliteit is geen onderdeel van dit onderzoek, voor zover het geen betrekking heeft op het beschrijven van de overlap en de implementatie van de theorie.

1.4 Doelstellingen

Het doel van dit onderzoek is het ontwikkelen en toetsen van een theorie om twee gemeenschappelijke entiteittypen te conciliëren zonder de aanwezigheid van gemeenschappelijke sterke sleutels. Om de bijbehorende onderzoeksvragen te kunnen beantwoorden, zijn de volgende doelstellingen opgesteld:

- Verkrijgen van informatie vergemakkelijken door onderzoek te doen naar algemeen inzetbare koppelingstechnieken, met als uiteindelijk doel het conciliëren van (de gemeenschappelijke entiteittypen in) informatiebronnen.
- Verduidelijken van de informatie door de definities van de informatiebronnen een belangrijke rol te laten spelen in:
    - het reconciliëren van objectgelijke entiteiten;
    - de definities van het gemeenschappelijk informatiemodel;
    - kwaliteitsbewaking.
- Bepalen van de kwaliteit van de conciliatie en de inhoud van de bronnen, met als doel:
    - inzicht krijgen in de onzekerheden tijdens en na het uitvoeren van de conciliatie;
    - de kwaliteit vaststellen van de oorspronkelijke informatiebronnen, zodat deze eventueel verbeterd kan worden;
    - de kwaliteit vaststellen van de gereconcilieerde informatie, zodat hier in analyses rekening mee gehouden kan worden.

De doelstellingen zullen, samen met de antwoorden op de onderzoeksvragen, besproken worden in hoofdstuk 7.

1.5 Aanpak

Een beproefde manier om objectgelijke entiteiten te reconciliëren is het gebruik maken van gemeenschappelijke attributen met semantische gelijkheid (gelijkheid qua betekenis). Op basis van definities wordt een analyse gemaakt van deze attributen. Daarna worden uitgebreide interviews gevoerd met bronexperts met als doel het verifiëren van de gevonden semantische overlap en het inventariseren van overige overlap.

Deze inventarisatie leidt uiteindelijk tot een verzameling expertkennis waarmee de gelijkheid tussen twee entiteiten beschreven kan worden. Uit deze expertkennis wordt een informatiemodel afgeleid dat bestaat uit de gemeenschappelijke entiteittypen.

Het theoretische deel van dit onderzoek beschrijft hoe gelijkheid op verschillende plaatsen in het informatiemodel bijdraagt aan de gelijkheid van één centraal gemeenschappelijk entiteittype: de objectgelijkheid. De objectgelijkheid wordt vervolgens gebruikt in de reconciliatie van objectgelijke entiteiten.

De theorie wordt getoetst door middel van een casus. Hiervoor is een prototype ontwikkeld waarin het mogelijk is de expertkennis en het gemeenschappelijk informatiemodel te gebruiken om entiteiten te reconciliëren. Bovendien slaat het prototype informatie op over de gemaakte conciliatie. De resultaten worden vervolgens getoetst op kwaliteit, zowel de kwaliteit van de conciliatie als de kwaliteit van de oorspronkelijke informatiebronnen.

1.6 Opbouw

De opbouw van deze scriptie is als volgt; in het volgende hoofdstuk wordt de onderzoeksomgeving en het theoretisch kader geschetst. Hoofdstuk 3 beschrijft de theorie die ontwikkeld is om objectgelijke entiteiten te reconciliëren ondanks beperkte overlap. Hoofdstuk 4 behandelt de casus die gebruikt is om de theorie in de praktijk te toetsen. Hoofdstuk 5 bespreekt EROS; het prototype waarmee de conciliatie van de casus is uitgevoerd. In hoofdstuk 6 worden de resultaten van de conciliatie geanalyseerd; zowel de kwaliteit als de correctheid van de gemaakte conciliatie. De probleemstelling en onderzoeksvragen worden besproken in hoofdstuk 7. Hierin zijn ook de conclusies en aanbevelingen opgenomen.

2 Achtergrond

Dit hoofdstuk beschrijft de onderzoeksomgeving – waarin het onderzoek is uitgevoerd – en het onderzoeksgebied – waarin het onderzoek zich afspeelt.

2.1 Onderzoeksomgeving

2.1.1 Organisatie

Het Wetenschappelijk Onderzoek- en Documentatiecentrum (WODC) is een toonaangevend wetenschappelijk kennisinstituut voor het brede veld van het Ministerie van Justitie, doordat zij producten en diensten aanbiedt inzake onderzoek, advisering en kennisverspreiding. Het WODC bestaat uit zeven afdelingen[1], waaronder vier onderzoeksafdelingen, één afdeling voor het uitbesteden van onderzoek en daarnaast nog een afdeling voor documentaire informatievoorziening. Dit onderzoek wordt uitgevoerd in de zevende afdeling, namelijk de afdeling Statistische Informatievoorziening en Beleidsanalyse (SIBa).

De afdeling SIBa heeft tot taak om op proactieve en reactieve wijze beleidsinformatie te leveren aan een brede en divers samengestelde klantengroep: bestuursdepartementen, uitvoerende diensten, politie, het landelijke Openbaar Ministerie, andere ministeries, de media, wetenschappelijke instellingen, enzovoort. Het gaat daarbij niet alleen om cijfers over de actuele stand van zaken binnen de verschillende justitiesectoren, maar ook om achtergrondgegevens waarmee een ontwikkeling in de juiste context kan worden beoordeeld. Daarnaast ondersteunt SIBa beleidsmakers bij het inschatten van effecten van voorgenomen beleidsmaatregelen. Hiervoor worden instrumenten (rekenmodellen) ontwikkeld waarmee bijvoorbeeld keteneffecten in beeld worden gebracht. Met deze instrumenten is het vervolgens mogelijk scenariostudies uit te voeren ter identificatie of doorrekening van globale trends en ter afweging van verschillende beleidsopties.

SIBa gebruikt voor haar onderzoek verschillende informatiebronnen, die ook op verschillende manieren worden aangeleverd. Intern wordt gebruik gemaakt van een Oracle database. In de documentbijlage Onderzoeksomgeving wordt verder ingegaan op de organisatie.

2.1.2 Data-analyse

In dit onderzoek zijn drie informatiebronnen gebruikt, die allemaal een deel van de strafrechtsketen betreffen:

- het Herkenningsdienstsysteem (HKS), registratie van processen-verbaal van aangifte van misdrijven;
- OMDATA, een informatiesysteem van het Parket-Generaal van het Openbaar Ministerie;
- de OBJD, een afgeleide van het Justitieel Documentatie Systeem (JDS) waarin strafbladen worden bijgehouden.

[1] Bron: www.wodc.nl, december 2005.

Deze paragraaf bespreekt allereerst de objecten die in de informatiebronnen voorkomen. Daarna worden de informatiebronnen zelf, alsmede hun onderlinge relatie, kort besproken. Bij de uitwerking van de casus (hoofdstuk 4) wordt verder ingegaan op het gebruik van de informatiebronnen.

Definities

Om de structuur van de informatiebronnen goed te kunnen begrijpen, is het van belang de objecten binnen de strafrechtsketen te kennen. Sommige objectdefinities hebben in de brondefinities verschillende synoniemen, welke hieronder ook genoemd worden.

Persoon
In dit onderzoek wordt gesproken over personen en persoonsreconciliatie. Binnen de strafrechtsketen kan een persoon verdachte (of dader) en/of slachtoffer zijn. In deze drie informatiebronnen staan personen als verdachte of dader centraal. De term *Persoon* verwijst dan ook naar deze rol van het object Persoon.

Proces-verbaal
Tegen een verdachte kunnen één of meer processen-verbaal (pv's) gemaakt worden. Een andere term voor proces-verbaal is *antecedent*. Een proces-verbaal, oftewel antecedent, kan bestaan uit één of meer misdrijven (delicten).

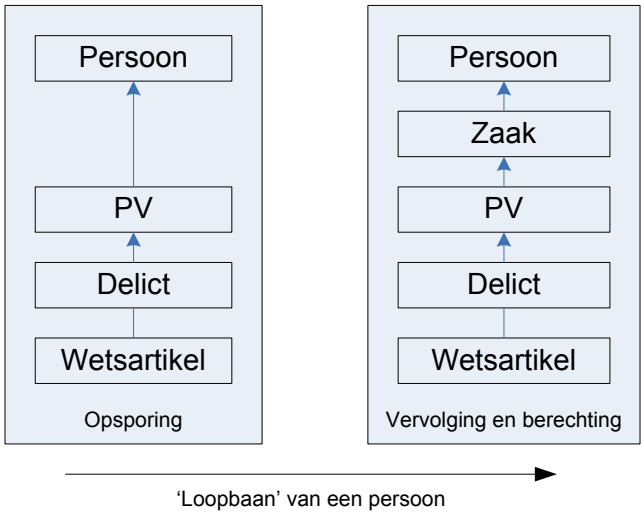
Zaak
Een (straf)zaak wordt bij het Openbaar Ministerie en de zittende magistratuur gedefinieerd als 'een proces-verbaal tegen één verdachte wegens één of meer strafbare feiten, dat bij het Openbaar Ministerie staat ingeschreven ter (verdere) afhandeling'. Eén verdachte (persoon) kan meer dan één strafbaar feit hebben gepleegd, terwijl anderzijds bij één strafbaar feit meerdere verdachten betrokken kunnen zijn. Zaken kunnen gekoppeld worden door een gezamenlijk parketnummer of parketnummerreeks. Doordat meerdere zaken samengevoegd kunnen worden, is het in de praktijk zo dat er meerdere processen-verbaal onder een zaak geregistreerd kunnen staan.

Delict
Een delict is een misdrijf gepleegd door een verdachte. Delicten worden ook wel *strafbare feiten* (kortweg feiten) genoemd. Een delict wordt beschreven door één of meer wetsartikelen. Verder worden delicten ingedeeld in rubrieken en sub-rubrieken volgens de CBS-standaard-classificatie.

Overig
Naast bovengenoemde objecten bevat elke informatiebron nog meer objecten, zoals zittingen in OMDATA. Deze objecten spelen, zoals later blijkt, geen rol in het onderzoek en worden daarom niet verder genoemd.

Overzicht

De relaties tussen bovengenoemde objecten zijn, gezien vanuit de strafrechtsketen, weergegeven in Figuur 2.1.



Figuur 2.1 – Relatie tussen de entiteittypen in de strafrechtsketen

In de loop van de strafrechtsketen komt het entiteittype Zaak erbij: een proces-verbaal wordt onder een zaak geplaatst zodra het door het Openbaar Ministerie in behandeling wordt genomen. Het HKS valt onder opsporing, OMDATA en OBJD onder vervolging en berechting.

In Bijlage A is meer achtergrondinformatie te vinden over de informatiebronnen. Voor de beeldvorming worden in deze paragraaf enkel de karakteristieken van de informatiebronnen getoond. De OBJD is uiteindelijk niet gebruikt in de conciliatie en wordt daarom niet genoemd. In de toetsing wordt een referentieset[1] gebruikt, waarvan ook de karakteristieken worden getoond.

Object	Aantal	Referentieset
Personen	932.064	10.000
Processen-verbaal	2.454.625	36.576
Delicten	10.248.943	76.440

Tabel 2.1 – Karakteristieken HKS

Object	Aantal	Referentieset
Personen	onbekend	8.705
Processen-verbaal	7.250.133	42.749
Delicten	8.498.824	106.308

Tabel 2.2 – Karakteristieken OMDATA

2.2 Onderzoeksgebied

Door middel van literatuuronderzoek is het wetenschappelijk gebied rond dit onderzoek in kaart gebracht. Een uitgebreid verslag hiervan is te vinden in Bijlage B. Deze paragraaf bespreekt beknopt het onderzoeksgebied.

Eén van de doelstellingen in dit onderzoek is het conciliëren van gemeenschappelijke entiteittypen. Om dit te kunnen doen moeten de schema's van de verschillende bronnen – ten minste voor wat betreft de gemeenschappelijke entiteittypen – worden geïntegreerd. Bij

[1] Een referentieset is een representatieve subset van de werkelijke informatiebronnen; in dit geval een combinatie van de informatiebronnen HKS, OMDATA en OBJD.

schema-integratie zijn twee gebieden te onderscheiden: schema-integratie *met* gemeenschappelijke sleutels ([AKWS95], [KCGS93], [DeM89], [LSS96]) en *zonder* gemeenschappelijke sleutels ([DSD02], [WM89]). Dit onderzoek vindt plaats in het laatste gebied.

Bij schema-integratie zonder gemeenschappelijke sleutels is het kernprobleem het identificeren en koppelen van entiteiten die naar hetzelfde object in de reële wereld verwijzen. Dit probleem staat bekend onder meerdere termen (*entity heterogeneity* [DSD98], *instance identification* [WM89], *merge/purge problem* [Gass85], *object isomerism* [CTK96], *common identifier problem* [HS95]); in dit onderzoek wordt de term entiteitreconciliatie gebruikt ([DSD02]) omdat deze term het beste aansluit op dit onderzoek. De eerder genoemde studies geven een oplossing voor dit probleem door de gemeenschappelijke entiteittypen afzonderlijk te bekijken. Dit onderzoek neemt een nieuwe invalshoek door een centraal gemeenschappelijk entiteittype te kiezen, waarin ook de omliggende entiteittypen bijdragen aan de conciliatie. Hierover is meer te lezen in paragraaf 3.2.

Wanneer er geen gemeenschappelijke sleutels voor handen zijn, moeten er andere attributen gebruikt worden voor de entiteitreconciliatie. Hiermee bevindt het onderzoek zich ook in het gebied van attribuutselectie. Bij entiteitreconciliatie wordt gezocht naar attributen met gelijke semantische betekenis of een andere mogelijkheid om de gelijkheid tussen twee entiteiten te bepalen. Dit kan automatisch gebeuren, op basis van definitie ([CER87], [BOT86], [TBDLW87]) of op basis van inhoud ([KCGS93], [DeM89]). Het kan ook gebeuren door bronexperts of door een combinatie van beide [DSD02]. Bij automatische attribuutselectie spelen vele problemen een rol ([DSD02], [Coh98], [ME96]), waar dit onderzoek zich niet op concentreert. Daarom wordt voor de beschrijving van gelijkheid gebruik gemaakt van expertkennis. De complexiteit van attribuutselectie speelt daarom niet in dit onderzoek: de attribuutselectie is onderdeel van de beschrijving van gelijkheid.

Bij het koppelen van informatiebronnen kan inconsistentie van attribuutwaarden een rol spelen. In dit onderzoek zeggen conflicterende attribuutwaarden iets over de kwaliteit van de informatiebronnen en/of de reconciliaties: één van de doelen van dit onderzoek. De informatie over conflicterende attribuutwaarden kan later gebruikt worden door in de bevraging van de geconcilieerde gegevens te werken met onzekerheid en onwetendheid. Choenni et al bespreken hoe deze informatie gebruikt kan worden in relationele informatiebronnen ([CBF04], [CBL06]).

Het identificeren van entiteiten kan een probleem zijn ([LSPR93], [PRSL93]). In dit onderzoek wordt aangenomen dat de te reconciliëren entiteiten per bron identificeerbaar zijn, d.w.z. door middel van een sterke sleutel terug te vinden zijn. Deze sleutel is niet gemeenschappelijk.

3 Theorie

Dit hoofdstuk bespreekt de theorie die ontwikkeld is om een antwoord te kunnen geven op de centrale probleemstelling, die betrekking heeft op entiteitreconciliatie ondanks beperkte overlap. De eerste paragraaf bespreekt hoe overlap gedefinieerd kan worden door gemeenschappelijke eigenschappen op attribuutniveau te beschrijven. De tweede paragraaf bespreekt hoe de gelijkheid gemodelleerd kan worden rond het gemeenschappelijke entiteittype dat geconcilieerd wordt. In de derde paragraaf wordt beschreven hoe uit de gelijkheid op attribuutniveau de objectgelijkheid tussen twee entiteiten van het gemeenschappelijk entiteittype wordt berekend. De reconciliatie van objectgelijke entiteiten wordt besproken in de vierde paragraaf. Paragraaf 3.5 geeft een korte bespreking van de theorie, gevolgd door een conclusie. Voor geïnteresseerden is de theorie in Bijlage H formeel uitgewerkt.

3.1 Beschrijving van gelijkheid

Objectgelijke entiteiten worden bij elkaar gezocht op basis van gelijkheid. Deze paragraaf laat zien dat gelijkheid beschreven kan worden aan de hand van de afstand tussen twee waarden die dezelfde eigenschap beschrijven. Deze beschrijving van gelijkheid berust op de werkelijkheid en is daarom onafhankelijk van de inhoud van de informatiebronnen.

3.1.1 Definitie van gelijkheid

In de wiskunde wordt gelijkheid tussen twee elementen, zoals entiteiten en attribuutwaarden, geschreven als een binaire relatie met twee uitkomsten: ongelijk (0) of gelijk (1). Ook in dit onderzoek worden elementen met elkaar vergeleken. Hierbij wordt beschreven in welke mate de twee elementen objectgelijk kunnen zijn (naar hetzelfde object verwijzen). Hoe meer de gemeenschappelijke informatie tussen twee elementen overlapt, hoe groter de mate van objectgelijkheid. De regels om objectgelijkheid te beschrijven worden vastgelegd in een definitie.

Definitie. Een gelijkheidswaarde beschrijft de mate van objectgelijkheid tussen twee elementen. De waarde ligt in het interval [0,1] en kan daarnaast ook de waarde *n.a.* (*not available*, niet beschikbaar) aannemen als de mate van objectgelijkheid niet bepaald kan worden.

Een gelijkheidswaarde van 0 betekent absolute ongelijkheid; de twee vergeleken elementen kunnen niet bij hetzelfde object horen. Een gelijkheidswaarde van 1 betekent maximale overlap in de elementen. Dit hoeft echter niet te betekenen dat de objecten gelijk zijn. Het is immers mogelijk dat de objecten nog door andere elementen beschreven worden. In sommige gevallen is het niet mogelijk om de objectgelijkheid tussen twee elementen te bepalen, bijvoorbeeld als één van de elementen informatie mist.

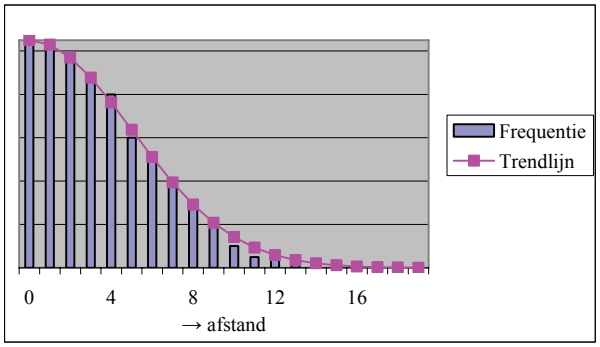
3.1.2 Gemeenschappelijke eigenschappen

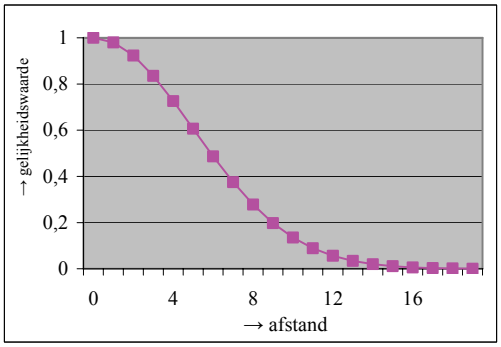
Gegeven twee informatiebronnen D_1 en D_2 en twee attributen $A_1 \in D_1$ en $A_2 \in D_2$ die gemeenschappelijke informatie bevatten. De attribuutwaarden van A_1 en A_2 worden genoteerd als a_1 en a_2 en zijn objectgelijk als ze tot objectgelijke entiteiten behoren. Als A_1 en A_2

semantisch gelijk zijn, dan levert de vergelijking van a_1 en a_2 – afgezien van representatie-verschillen, etc. – een gelijkheidswaarde van 0 of 1 op als respectievelijk $a_1=a_2$ of $a_1\neq a_2$. Als A_1 en A_2 niet semantisch gelijk zijn, dan kan er toch een bepaald verband bestaan. Dit verband wordt gedefinieerd door middel van afstand.

Afhankelijk van de attributen kan de afstand op verschillende manieren berekend worden ([DSD02]). Voor numerieke attributen (inclusief datums) kan het verschil bijvoorbeeld berekend worden door aftrekking. Als A_1 en A_2 één gemeenschappelijke eigenschap beschrijven, dan concentreert de afstand tussen twee objectgelijke waarden zich rond één verwachte waarde (d_0). Zo geldt voor semantisch gelijke attributen: $d_0=0$. Voor semantisch ongelijke attributen moet een afstand berekend worden. Deze afstand hoeft niet altijd exact gelijk zijn aan de verwachte waarde, maar zal hier wel in de buurt liggen. Dit gedrag kan beschreven worden door een afstandsverdeling. Een afstand in de afstandsverdeling wordt vervolgens door middel van een gelijkheidsfunctie omgezet in een gelijkheidswaarde.

Voorbeeld. Gegeven een gemeenschappelijke eigenschap van een misdrijf die het verband tussen de pleegdatum en de datum waarop het proces-verbaal is opgemaakt, beschrijft. De meeste processen-verbaal worden op dezelfde dag opgemaakt, er geldt: $d_0=0$. Figuur 3.1 en Figuur 3.2 tonen respectievelijk de verwachte afstandsverdeling en de bijbehorende gelijkheidsfunctie.





Figuur 3.1 (l) – Afstandsverdeling en frequentiefunctie
Figuur 3.2 (r) – Gelijkheidsfunctie

Definitie. Een afstandsverdeling is een histogram van de verwachte frequenties van afstanden tussen objectgelijke waarden: de frequenties van de afstanden worden uitgezet tegen de afstanden zelf.

Definitie. De frequentiefunctie *freq*(d) benadert de afstandsverdeling voor een afstand d. Het maximum van de frequentiefunctie is de frequentie van de verwachte afstand *freq*(d_0).

Definitie. De gelijkheidsfunctie *sim*(a_1, a_2) berekent de gelijkheidswaarde van een gemeenschappelijke eigenschap en wordt gedefinieerd als:

$$sim(a_1, a_2) = \frac{freq(\delta(a_1, a_2))}{freq(d_0)}$$

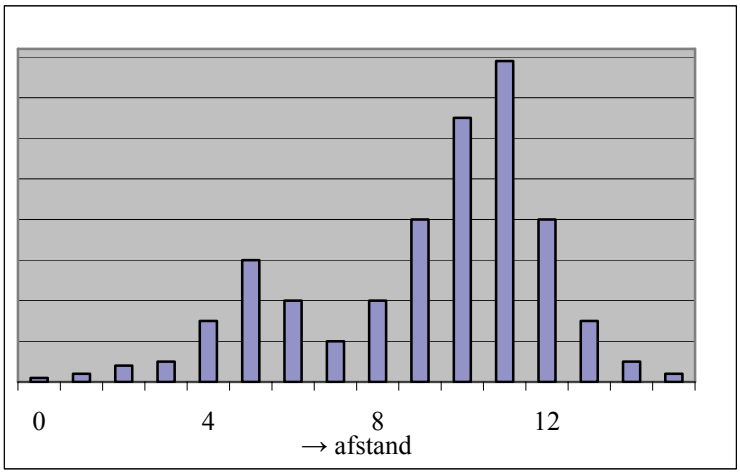
Doordat de frequentie gedeeld wordt door het maximum van de frequentiefunctie, heeft de gelijkheidsfunctie een bereik van 0 tot en met 1. Als de afstand $\delta(a_1, a_2)$ niet bepaald kan worden, dan heeft de gelijkheidsfunctie als uitkomst *n.a.*

Voor sterke eigenschappen levert de gelijkheidsfunctie een duidelijke scheiding tussen objectgelijke waarden en waarden die dit niet zijn. Hoe zwakker de eigenschappen, hoe meer interferentie er optreedt van objectongelijke waarden. Door het instellen van een betrouwbaarheidsinterval zou een grens bepaald kunnen worden. Dit wordt overgelaten aan toekomstig onderzoek.

Meerdere afstandsverdelingen

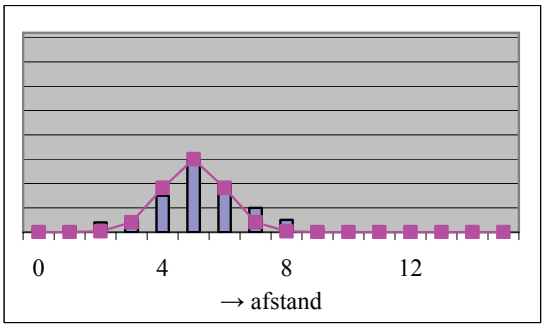
Tot nu toe is uitgegaan van attributen die slechts één gemeenschappelijke eigenschap beschrijven. Attributen met meer gemeenschappelijke eigenschappen kunnen echter ook beschreven worden, als voor elke attribuutwaarde bepaald kan worden welke eigenschap beschreven wordt. Zodoende zijn verschillende gelijkheidsfuncties te definiëren voor elke gemeenschappelijke eigenschap, waarbij de domeinen van de gelijkheidsfuncties paarsgewijs disjunct (d.w.z. geen gemeenschappelijke elementen hebben) zijn.

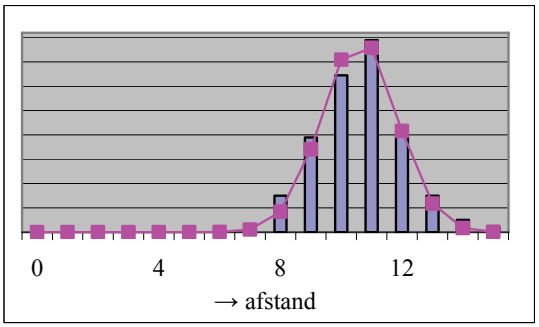
Voorbeeld. Gegeven een attribuut met twee gemeenschappelijke eigenschappen waarbij de pleegdatum van een misdrijf wordt vergeleken met de datum van het eindvonnis. Stel, er zijn twee soorten vonnissen; een vonnis voor snelrecht en een vonnis voor andere zaken. De verwachte afstand tussen pleegdatum en eindvonnis is voor snelrechtzaken veel kleiner dan voor andere zaken. Stel dat het histogram er als volgt uit ziet (de afstanden zijn weken):



Figuur 3.3 – Afstandsverdeling voor snelrecht en overige zaken samen

Het histogram heeft twee lokale maxima op de afstanden 5 en 11; voor elke gemeenschappelijke eigenschap één. Door de eigenschappen te scheiden, kan voor allebei een afstandsverdeling worden vastgesteld. Elke afstandsverdeling heeft zijn eigen gelijkheidsfunctie, waardoor de piek in Figuur 3.4 nu ook maximale gelijkheid oplevert.





Figuur 3.4 (l) – Afstandsverdeling en frequentiefunctie voor snelrecht (d_0=5)
Figuur 3.5 (r) – Afstandsverdeling en frequentiefunctie voor overige zaken (d_0=11)

Ruis

In sommige gevallen kan de afstand tussen twee objectgelijke waarden extreem afwijken van de verwachting. Dit soort gevallen wordt *ruis* genoemd. In de frequentiefunctie krijgen deze gevallen zo'n lage waarde, dat de gelijkheid nagenoeg nul is. Als het belangrijk is om de ruis op te vangen, dan moet de gelijkheid voor deze gevallen verhoogd worden. Ruis is echter niet te onderscheiden van reële afstanden. Daarom worden alle waarden in de omgeving verhoogd tot aan de zogenaamde *ruisdrempel*. De ruisdrempel is een zwaar middel, omdat ook het vergelijken van objectongelijke waarden hiermee meer gelijkheid oplevert. De ruisdrempel moet daarom alleen worden ingezet voor de afstanden waar de ruis verwacht wordt.

Voorbeeld. Gegeven het voorbeeld bij Figuur 3.1. Stel dat 90% van de processen-verbaal binnen een week wordt opgemaakt, van de overige 10% is het onbekend. De enige zekerheid is volgens bronexperts dat een proces-verbaal binnen een jaar wordt opgemaakt. Van de 90% kan een betrouwbare afstandsverdeling worden gemaakt. De laatste 10% is op een onbekende manier verspreid over een jaar. Om deze 10% niet bij voorbaat uit te sluiten, kan gebruik worden gemaakt van een ruisdrempel. De ruisdrempel verhoogt de gelijkheid bij een afstand tussen 7 dagen en een jaar net genoeg om meegeteld te worden.

3.1.3 Expertkennis

De expertkennis wordt geformaliseerd in kennisregels. Elke kennisregel beschrijft hoe de gelijkheid tussen twee attributen berekend kan worden en levert gelijkheidswaarde op. Aangezien een attribuut meerdere eigenschappen kan beschrijven, kan een kennisregel opgebouwd zijn uit meerdere gelijkheidsfuncties.

Definitie. Gegeven twee attribuutwaarden a_1 en a_2 die n gemeenschappelijke eigenschappen beschrijven. De domeinen van de attributen zijn respectievelijk D_1 en D_2. De paarsgewijs disjuncte subdomeinen van elke gemeenschappelijke eigenschap i zijn respectievelijk $D_{1,i}$ en $D_{2,i}$. Een gelijkheid uit een kennisregel wordt nu gedefinieerd als:

$$sim_{rule} = \begin{cases} sim_1(a_1, a_2) & \text{als } a_1 \in D_{1,1} \wedge a_2 \in D_{2,1} \\ sim_2(a_1, a_2) & \text{als } a_1 \in D_{1,2} \wedge a_2 \in D_{2,2} \\ \vdots & \vdots \\ sim_n(a_1, a_2) & \text{als } a_1 \in D_{1,n} \wedge a_2 \in D_{2,n} \\ n.a. & \text{anders} \end{cases}$$

Niet elke kennisregel is even waardevol in de bepaling van gelijkheid. Sterk discriminerende attributen zijn waardevoller in de bepaling dan zwak discriminerende attributen.

Voorbeeld. Gegeven twee sets van persoonsentiteiten – elk uit een andere informatiebron – die vergeleken worden:

Bron I			Bron II		
id	**geb.datum**	**geslacht**	**id**	**geb.datum**	**geslacht**
1	01-01-1975	m	a	01-01-1975	m
2	01-03-1980	m	b	01-03-1980	m
3	23-04-1977	m	c	23-04-1977	m
4	01-01-1975	v	d	01-01-1975	v
5	15-02-1965	v	e	15-02-1965	v
6	17-07-1974	v	f	17-07-1974	v

Tabel 3.1 – Sets van persoonsentiteiten

Per attribuutwaarde wordt gekeken naar het aantal entiteitcombinaties dat mogelijk objectgelijk is:

geb.datum	mogelijkheden
15-02-1965	1
17-07-1974	1
01-01-1975	4
23-04-1977	1
01-03-1980	1

geslacht	mogelijkheden
m	9
v	9

Tabel 3.2 – Mogelijkheden per attribuut

Duidelijk is dat geboortedatum beter in staat is het aantal mogelijkheden te beperken dan geslacht. De kennisregel die de gemeenschappelijk eigenschap 'geboortedatum' beschrijft, krijgt daarom meer gewicht. Het bepalen van het gewicht van kennisregels valt onder expertkennis, maar kan automatisch berekend worden door de mate van discriminatie om te zetten in een gewicht. In Bijlage C is dit verder uitgewerkt.

Verder kunnen bronexperts ook ruis beschrijven aan de hand van twee variabelen: de ruisdrempel en het afstandsbereik waarover deze drempel geldt.

3.2 Modellering van gelijkheid

Deze paragraaf bespreekt de distributie van gelijkheid. De paragraaf laat zien dat het reconciliëren van objectgelijke entiteiten van één gemeenschappelijk entiteittype het meest efficiënt kan met één of meer hiërarchische informatiemodellen.

3.2.1 Attributen en knopen

Objectgelijke entiteiten worden bepaald door de gemeenschappelijke informatie tussen twee bronnen te vergelijken. De gemeenschappelijke informatie wordt beschreven door gemeenschappelijke eigenschappen, zoals besproken in paragraaf 3.1. Deze eigenschappen worden gerepresenteerd door attributen.

In een informatiebron kunnen meerdere objecten voorkomen. Dit onderzoek concentreert zich op objecten die in beide informatiebronnen voorkomen en dus een gemeenschappelijk entiteittype hebben. Door attributen te plaatsen onder deze objecten kunnen ze vergeleken worden. Een gemeenschappelijk entiteittype is immers een knooppunt tussen de informatiebronnen, waardoor de gemeenschappelijke eigenschap zich manifesteert in objectgelijke entiteiten.

Definitie. *Objectgelijke attributen* zijn attributen uit verschillende bronnen die één of meerdere gemeenschappelijke eigenschap(pen) representeren, waaruit blijkt dat ze naar hetzelfde object verwijzen.
Definitie. Een *knoop* is een gemeenschappelijk entiteittype in twee informatiebronnen.
Definitie. Een *centrumknoop* is de knoop waarvoor conciliatie gewenst is.

Objectgelijke attributen beschrijven een gemeenschappelijke eigenschap[1] en zullen daarom meestal behoren tot een gemeenschappelijk entiteittype. Als de attributen geen gemeenschappelijk entiteittype hebben of verschillende entiteittypen hebben[2], dan hoort de gemeenschappelijke eigenschap wel degelijk bij één gemeenschappelijk entiteittype. Het attribuut wordt dan onder de bijbehorende knoop geplaatst. Het plaatsen van objectgelijke attributen bij de bijbehorende knoop is nodig voor een goede vergelijking.

Voorbeeld. Personen worden vergeleken met behulp van twee delicteigenschappen: pleegdatum en delictsoort. Gegeven twee objectongelijke persoonsentiteiten uit verschillende bronnen:

Entiteiten uit bron I		**Entiteiten uit bron II**	
Pleegdatum	**Delictsoort**	**Pleegdatum**	**Delictsoort**
12-3-1980	geweld	12-3-1980	diefstal
5-11-1995	diefstal	5-11-1995	geweld

Tabel 3.3 - Objectongelijke persoonsentiteiten uit verschillende bronnen

Als de attributen onder Persoon geplaatst worden, dan zijn ze meerwaardig. Voor beide entiteiten geldt dan: pleegdatum={12-3-1980, 5-11-1995} en delictsoort={geweld, diefstal}. De persoonsentiteiten zouden gelijk zijn. Door de attributen onder Delict te plaatsen, worden de delicten onder persoon vergeleken. Het is duidelijk dat de delicten onder de entiteit uit bron I niet gelijk zijn aan de delicten onder de entiteit uit bron II. Als gevolg hiervan zijn de persoonsentiteiten ook objectongelijk.

3.2.2 Distributie van gelijkheid

De objectgelijke attributen zijn nu toegekend aan knopen. Deze knopen hebben een relatie met de centrumknoop, anders kunnen de attributen het centrale object niet beschrijven. De relaties zijn binair; relaties met een hogere graad worden later in deze paragraaf besproken. Uiteindelijk moeten de gelijkheidswaarden uit attribuutvergelijkingen bij de centrumknoop komen. Het doorgeven ('distributie') van gelijkheid wordt beschreven als een gerichte graaf $G=(V,E,c)$, met de verzameling van knopen V, de verzameling van pijlen $E = \{(v,c) \mid v \in V \wedge v \neq c\}$ en de centrumknoop c. De graaf die de distributie van gelijkheid beschrijft, wordt de *gelijkheids-distributie* genoemd. Later in deze paragraaf wordt een exacte definitie van dit begrip gegeven.

De gelijkheidsdistributie zoals hierboven beschreven, is een ster met pijlen naar het middelpunt: de centrumknoop. Knopen onderling kunnen echter ook een relatie hebben. Elke knoop is immers een object op zich, dat mede beschreven kan worden door andere knopen. Aangezien elke knoop mede de gelijkheid van de centrumknoop beschrijft, is het wenselijk elke knoop zo goed mogelijk te beschrijven. In de gelijkheidsdistributie worden gerichte lijnen geplaatst tussen knopen met een directe relatie.

Voorbeeld. Stel, er is een gelijkheidsdistributie $G(V,E,c)$ met $V=\{A,B,C\}$, $E=\{AC,BC\}$ en $c=C$. Stel nu dat de relatie AB ook bestaat. Dan is de opname van de relatie AC gebaseerd op de transitieve afsluiting $AB \wedge BC \rightarrow AC$. In dit geval moet niet AC, maar AB worden opgenomen:

[1] Bij meerdere eigenschappen, die naar verschillende entiteittypen verwijzen, kan het attribuut worden geplaatst onder meerdere entiteittypen. Alleen de eigenschappen die het entiteittype beschrijven worden dan vergeleken.

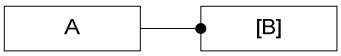
[2] 'Geboortedatum is kleiner dan Pleegdatum' beschrijft een delicteigenschap.

$E'=\{AB,BC\}$. Een concreet voorbeeld: een persoon (C) wordt beschreven door zijn processen-verbaal (B) en de delicten op deze processen-verbaal (A). De delicten beschrijven de processen-verbaal (AB) en daarmee uiteindelijk ook persoon (BC), maar ze beschrijven persoon niet rechtstreeks (AC). Door de gelijkheid van delicten door te geven via hun proces-verbaal, wordt het proces-verbaal ook beter beschreven.

Als de gelijkheidsdistributie een lus of cykel bevat, dan beschrijft een knoop zichzelf of beschrijven twee knopen *elkaar*. Dit is niet mogelijk in één gelijkheidsdistributie; de oplossing hiervoor wordt later in deze paragraaf gegeven. De gelijkheid in de gelijkheidsdistributie wordt in de richting van de centrumknoop doorgegeven. Er is daarom sprake van een georiënteerde graaf (een gerichte graaf zonder lussen). Verder is de gelijkheidsdistributie samenhangend, omdat elke knoop een pad heeft naar de centrumknoop. De gelijkheidsdistributie kan nu als volgt gedefinieerd worden.

Definitie. Een gelijkheidsdistributie $G(V,E,c)$ is een georiënteerde samenhangende graaf, waarin $\forall v \in V : v \sim *c$ (elke knoop heeft een pad naar de centrumknoop c).

Een pijl in de gelijkheidsdistributie betekent dat de gelijkheid van een knoop naar een andere knoop moet worden doorgegeven. De gelijkheid in een knoop kan pas berekend worden als de gerelateerde knopen achter alle inkomende pijlen zijn berekend, d.w.z. de gelijkheid van de andere knoop bekend is. In een grafische representatie worden pijlen vervangen door bollen om verwarring met de notatie in datamodellen te voorkomen. Verder wordt de centrumknoop weergegeven tussen blokhaken.



Figuur 3.6 – Knoop A distribueert gelijkheid naar centrumknoop B

Binnen de gelijkheidsdistributie worden twee soorten gelijkheid gedefinieerd.

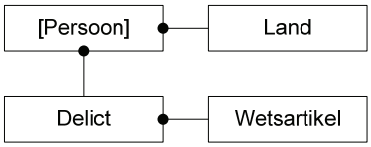
Definitie. De entiteitgelijkheid s_A beschrijft de gelijkheid binnen een knoop A.
Definitie. De objectgelijkheid S(A) is de formule voor de totale gelijkheid van A.

De formule voor objectgelijkheid S(A) bestaat uit de entiteitgelijkheid van A en de objectgelijkheid van knopen achter inkomende pijlen naar A. Binnen de formule S(A) wordt de operator $\circ$ gebruikt, die de verschillende gelijkheidswaarden combineert. De operator en entiteitgelijkheid wordt uitgewerkt in §3.3. Voor nu gedragen zij zich respectievelijk als een vermenigvuldiger en constante.

Voorbeeld. Gegeven de gelijkheidsdistributie $G(V,E,c)$ met $V=\{A,B\}$, $E=\{AB\}$ en $c=B$. Er geldt: S(A)=s_A en S(B)= S(A) $\circ s_B$.

Eén van de eigenschappen die volgt uit de definitie van de gelijkheidsdistributie is dat er altijd minstens één knoop is met enkel uitgaande pijlen. Door dit gegeven kan de gelijkheid in elke knoop berekend worden met het volgende algoritme.

1. Bereken de objectgelijkheid van knopen met ingraad 0. Voor deze knopen is de objectgelijkheid gelijk aan de entiteitgelijkheid.
2. Distribueer de objectgelijkheid over de uitgaande pijlen naar gerelateerde knopen.
3. Bereken de objectgelijkheid van de gerelateerde knopen, waarvoor alle inkomende objectgelijkheid berekend is.
4. Herhaal stap 2 en 3 totdat de objectgelijkheid in elke knoop berekend is.



Figuur 3.7 – Voorbeeld van een gelijkheidsdistributie

Voorbeeld. Gegeven de gelijkheidsdistributie uit Figuur 3.7. In stap 1 worden Land en Wetsartikel berekend. In stap 2 wordt de objectgelijkheid van Land en Wetsartikel gedistribueerd naar respectievelijk Persoon en Delict. In stap 3 wordt Delict berekend. Vervolgens wordt stap 2 herhaald door de objectgelijkheid van Delict te distribueren naar Persoon en wordt in stap 3 de objectgelijkheid van Persoon berekend.

De knopen in het voorbeeld worden in een bepaalde volgorde berekend, die afhankelijk is van de gelijkheidsdistributie. Als laatste wordt de objectgelijkheid van de centrumknoop berekend. De volgorde waarin objectgelijkheid wordt berekend, wordt weergegeven in een gelijkheidsstelsel.

Definitie. Het gelijkheidsstelsel S bevat de formules voor objectgelijkheid in de volgorde waarin ze opgelost kunnen worden. Onderaan staat de objectgelijkheid voor de centrumknoop.

Voorbeeld. Gegeven de gelijkheidsdistributie G uit het vorige voorbeeld. Er geldt:

$$S = \begin{cases} S(W) = s_W \\ S(L) = s_L \\ S(D) = s_D \circ S(W) \\ S(P) = S(L) \circ s_P \circ S(D) \end{cases}$$

Door distributie van gelijkheid van boven naar beneden geldt uiteindelijk $S(P)=s_L \circ s_P \circ s_D$.

In een gelijkheidsdistributie kunnen lussen, cykels en relaties met een hogere graad dan binair (zoals ternair) niet worden weergegeven. Toch kan voor deze gevallen een gelijkheidsstelsel worden opgesteld, waarmee de objectgelijkheid effectief berekend wordt.

Een cykel in de gelijkheidsdistributie betekent dat knopen elkaar, eventueel via andere knopen, beschrijven. Een cykel kan opgelost worden door een lijn weg te halen. Als de knopen daarna te slecht beschreven worden, dan is de cykel onmisbaar. Een (onmisbare) cykel kan beschreven worden in twee formules door de recursie – die ontstaat door elkaar beschrijvende knopen – te verwijderen, zoals het volgende voorbeeld laat zien.

Voorbeeld. Gegeven een gelijkheidsdistributie $G=(V,E,c)$ met $V=\{A,B\}$, $E=\{AB,BA\}$ en $c=B$. Er geldt: $S(A)=s_A \circ S(B)$ en $S(B)=s_B \circ S(A)$. Dit stelsel is onoplosbaar. Door eerst A te berekenen met alleen de entiteitgelijkheid uit B is het stelsel wel oplosbaar: $S(A)=s_A \circ s_B$ en $S(B)=s_B \circ S(A)$.

Relaties met een hogere graad dan binair worden beschreven aan de hand van een ternaire relatie. Stel, de knopen A, B en C hebben een ternaire relatie. Deze knopen beschrijven elkaar, anders is het niet nodig de relatie in de gelijkheidsberekening op te nemen. In een ternaire relatie wordt elke knoop beschreven door twee andere knopen. Er geldt: $S(A)=s_A \circ S(B) \circ S(C)$, $S(B)=S(A) \circ s_B \circ S(C)$ en $S(C)=S(A) \circ S(B) \circ s_C$. De knopen vormen een cykel. Door deze cykel op te lossen, vervalt – in de berekening – de (ternaire) relatie tussen de knopen.

Meer gevallen uit de E/R modellering worden hier niet besproken, zoals relaties met attributen, generalisaties en specialisaties. Al deze gevallen kunnen, om in deze theorie te passen, worden omgezet naar entiteittypen door elk 'record' een uniek getal toe te kennen. Zo zijn ze geschikt om als knopen gebruikt te worden.

3.2.3 Informatiemodel

In een gelijkheidsstelsel zoals beschreven in de vorige paragraaf wordt de gelijkheid knoop voor knoop berekend. Hierbij worden alle mogelijke combinaties berekend, terwijl dit wellicht niet nodig is. Omdat in dit onderzoek gewerkt wordt met grote informatiebronnen, wordt gezocht naar een manier om zo min mogelijk combinaties door te rekenen.

Voorbeeld. Gegeven de gelijkheidsdistributie uit Figuur 3.7, alleen zonder de knoop Wetsartikel. Het algoritme uit de vorige paragraaf berekent eerst Land (*L*) en Delict (*D*) en vervolgens Persoon (*P*). Er zijn echter veel minder landen dan personen, en veel minder personen dan delicten. Efficiënter zou dan ook zijn om alleen persooncombinaties te berekenen waarvoor de landcombinatie niet ongelijk (d.w.z. groter dan nul of *n.a.*) is en alleen delict-combinaties te berekenen waarvoor de persooncombinatie (tot dan toe) niet ongelijk is.

De distributie van gelijkheid is in bovenstaand voorbeeld te schrijven als een graaf $G(V,E,c)$ met $V=\{L,P,D\}$, $E=\{LP,DP\}$ en $c=P$. Daarnaast zegt het voorbeeld iets over de relaties tussen de knopen zelf. Als een landcombinatie ongelijk is, dan kunnen persooncombinaties met deze landcombinatie niet gelijk zijn. Deze persooncombinaties hoeven daarom niet eens vergeleken te worden.

Tabel 3.4 toont mogelijke vergelijkingen. Hiervoor worden de formules voor objectgelijkheid uitgebreid met entiteitidentificaties. Zo levert de formule $S(L)(1,2)$ de objectgelijkheid tussen landsentiteiten uit bron I en II met respectievelijk een identificatie van 1 en 2. De entiteitcombinaties en hun gelijkheidswaarden zijn voorbeelden.

Landcombinaties	**Persooncombinaties**	**Delictcombinaties**
$S(L)(1,1)=1$	$S(P)(1,1)=1$	$S(D)(1,1)=1$ ...
	$S(P)(2,2)=0,5$	$S(D)(2,2)=0$ $S(D)(2,3)=0,5$ $S(D)(3,2)=0,5$ $S(D)(3,3)=1$ ...
	...	...
$S(L)(1,2)=0,8$	$S(P)(1,3)=1$	$S(D)(1,4)=0,6$ ...
	$S(P)(2,4)=0$	- (altijd ongelijk)
	...	...
$S(L)(1,3)=0$	- (altijd ongelijk)	- (altijd ongelijk)
...	...	...

Tabel 3.4 – Verschillende combinaties

Persoonscombinaties worden alleen berekend als de landcombinatie niet ongelijk is. Evenzo worden delictcombinaties alleen berekend als de persoonscombinatie berekend én niet ongelijk is. Het aantal vergelijkingen wordt hiermee verlaagd. Hoe meer vergelijkingen in een begin stadium ongelijkheid opleveren, hoe efficiënter de gelijkheidsberekening. In dit voorbeeld kan

het aantal persoonsvergelijkingen verlaagd worden door clustering op landcombinaties. In deze gevallen wordt Persoon de kindknoop en Land de ouderknoop genoemd.

Definitie. In een één-op-meer relatie tussen twee knopen wordt de eerste knoop de *ouderknoop* en de tweede knoop de *kindknoop* genoemd (één ouder heeft meerdere kinderen).

De clustering die in deze theorie gebruikt wordt, maakt gebruik van één ouderknoop. Dit is een keuze die wordt toegelicht in Bijlage C. Kort samengevat wordt hierin besproken dat het clusteren met slechts één ouder vaak voldoende is, omdat maximale efficiëntie behaald wordt door gedeeltelijke clustering. Bovendien komen andere vormen van clustering uiteindelijk ook neer op clustering met één ouder.

De knopen in het voorbeeld vormen nu een hiërarchische structuur: Land (L) is ouder van Persoon (P) en Persoon is op zijn beurt weer ouder van Delict (D). Verder is het gelijkheidsstelsel van de gelijkheidsdistributie: $S(L)=s_L$, $S(D)=s_D$ en $S(P)=S(L)\circ s_P\circ S(D)$. In de formule voor persoonsgelijkheid $S(P)$ zijn de variabelen gesorteerd volgens de hiërarchische structuur. Te zien is dat hierdoor de beoogde clustering ontstaat: als $S(L)$ nul is, dan hoeft $s_P\circ S(D)$ niet meer berekend te worden: de uitkomst van de formule is altijd nul.

Voordat de clustering algemeen gedefinieerd wordt, wordt voor de leesbaarheid een verkorte notatie geïntroduceerd voor de variabelen in de formules; $S(A)$ wordt geschreven als $\underline{a}$ en de constante s_A als a. Verder wordt de operator weggelaten, zoals bij vermenigvuldiging vaker gebruikelijk is.

De formule voor objectgelijkheid in een centrumknoop (c) bestaat in het algemeen uit maximaal één ouderknoop (o), de entiteitgelijkheid en kindknopen ($k_1..k_n$). De ouderknoop kan op zijn beurt weer een ouderknoop (p) en kindknopen ($l_1..l_m$) hebben. Eén van de kinderen is de centrumknoop, stel $l_m=c$. Eerst wordt een eenvoudige geval besproken met $m=2$ en $n=1$. Het gelijkheidsstelsel S is dan:

$$S=\begin{cases}\underline{p}=p\\ \underline{l_1}=l_1\\ \underline{o}=\underline{p}o\underline{l_1}\\ \underline{k_1}=k_1\\ \underline{c}=\underline{o}c\underline{k_1}\end{cases}$$

Het gelijkheidsstelsel S kan geschreven worden in één formule $\underline{c}$. Hiervoor worden alle formules ingevuld in $\underline{c}$, waarbij ingevulde formules tussen haakjes geplaatst worden:

$S=\underline{c}=((p)o(l_1))c(k_1)$

De clustering is van links naar rechts af te lezen: als de entiteitgelijkheid uit p nul is, dan hoeft c niet berekend te worden, etc. Door het gebruik van de haakjes zijn bovendien de ouder/kind relaties af te lezen: voor $(a)b(c)$ geldt dat a een ouderknoop en c een kindknoop van b is. Het algemene gelijkheidsstelsel S wordt nu:

$S=\underline{c}=((p)o(l_1..l_{m\text{-}1}))c(k_1..k_n)$

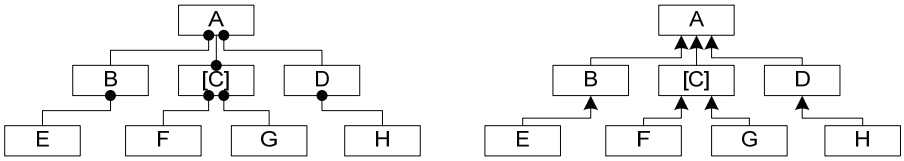
Een formule voor objectgelijkheid wordt grafisch weergegeven door middel van een informatiemodel.

Definitie. Een informatiemodel is een hiërarchisch datamodel van de knopen in een gelijkheidsdistributie met hun ouder/kind relaties, waarmee een formule voor objectgelijkheid

wordt weergegeven. In het informatiemodel worden soms ook de gebruikte attributen weergegeven.

Het informatiemodel is een *hiërarchisch datamodel* [SKS05] en heeft dus een boomstructuur. De wortel van de boom is gelijk aan de knoop die aan het begin van de formule staat.

Voorbeeld. Gegeven een gelijkheidsdistributie $G(V,E,c)$ met $V=\{A,B,C,D,E,F,G,H\}$, $E=\{AC,BA,DA,EB,FC,GC,HD\}$ en $c=C$. De gelijkheidsdistributie en clustering wordt beschreven door $\underline{c} = a(b(e)d(h))(c(fg))$.



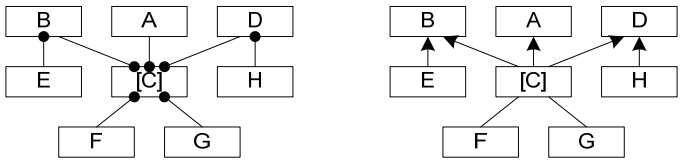
Figuur 3.8 – Gelijkheidsdistributie (l) en bijbehorend informatiemodel (r)

De gelijkheidsdistributie wordt in dezelfde boomstructuur weergegeven als het informatiemodel. Op die manier kan aan de structuur van de gelijkheidsdistributie het bijbehorende informatiemodel worden afgelezen (zie Figuur 3.8).

Uit voorgaande tekst blijkt dat de gewenste clustering alleen mogelijk is in een hiërarchische boomstructuur. Met andere woorden: de gelijkheidsdistributie kan alleen gebruik maken van clustering als het bijbehorende informatiemodel een hiërarchisch datamodel is. Het hiërarchisch datamodel legt twee beperkingen op aan de gelijkheidsdistributie:

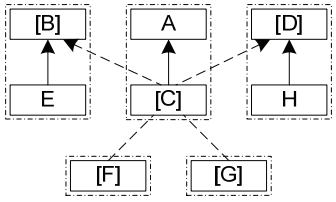
- meer-op-meer relaties zijn niet toegestaan
- een knoop mag slechts één ouderknoop hebben

Voorbeeld. Gegeven een gelijkheidsdistributie en het bijbehorende datamodel, met meer-op-meer relaties en een centrumknoop met meerdere ouderknopen:



Figuur 3.9 – Gelijkheidsdistributie (l) en bijbehorend datamodel (r)

In deze gelijkheidsdistributie is clustering niet mogelijk. Stel, de gebruiker kiest uit de ouderknopen A, B en D de knoop A als ouderknoop voor clustering. De knopen B en D moeten dan vooraf bekend zijn. Ook de knopen F en G kunnen – vanwege de meer-op-meer relatie – niet geplaatst worden in één hiërarchisch informatiemodel; er ontstaan meerdere informatiemodellen, zoals weergegeven in Figuur 3.10. Hierin zijn de informatiemodellen omrand. De virtuele relaties tussen knopen in verschillende modellen zijn gestippeld.



Figuur 3.10 – Meerdere hiërarchische informatiemodellen

Deze procedure kan herhaald worden voor elk ontstaan datamodel, totdat alle datamodellen hiërarchisch zijn. Als de hiërarchische datamodellen vervolgens in omgekeerde volgorde worden berekend, dan zijn alle gelijkheidswaarden op het gewenste moment berekend.

Bovendien wordt zoveel mogelijk gebruik gemaakt van clustering. Er ontstaat een gelijkheidsstelsel, waarin elke formule de objectgelijkheid van een centrumknoop in een hiërarchisch model berekent. Bij Figuur 3.10 hoort bijvoorbeeld het volgende gelijkheidsstelsel *S* (in verkorte notatie).

$$S = \begin{cases} \underline{b} = b(e) \\ \underline{d} = d(h) \\ \underline{f} = f \\ \underline{g} = g \\ \underline{c} = (a)\underline{b}\underline{d}\,\underline{f}\,\underline{g}c \end{cases}$$

De gestippelde relaties in Figuur 3.10 vormen een link tussen de verschillende informatiemodellen en geven aan dat de gelijkheid uit een ander informatiemodel afkomstig is. Deze relaties worden *virtuele relaties* genoemd, analoog aan de virtuele records die in een hiërarchische database model gebruikt worden. Virtuele records bevatten een verwijzing naar het fysieke record, dat ergens anders is opgeslagen [SKS05]. Virtuele relaties bevatten ook slechts een verwijzing. Deze verwijzing wordt gebruikt om de gelijkheid uit het andere informatiemodel op te halen.

Virtuele relaties – zoals in het gelijkheidsstelsel *S* bij Figuur 3.10 – worden voor de knoop geplaatst waarin ze nodig zijn, zonder gebruik van haakjes: het zijn immers geen ouder- of kindknopen. De gelijkheidswaarden uit virtuele records hoeven alleen maar opgehaald te worden uit het andere model en zijn dus snel te 'berekenen'. In de clustering worden ze daarom nog voor de berekening van de knoop zelf geplaatst.

3.3 Berekening van gelijkheid

Deze paragraaf laat zien hoe de gelijkheid wordt berekend in een informatiemodel. Bij het rekenen met gelijkheid wordt gebruik gemaakt van een gelijkheidsoperator (∘). De gelijkheidsoperator wordt gebruikt om uit verschillende gelijkheidswaarden één waarde te berekenen. Een gelijkheidswaarde ligt in het interval [0,1], maar kan ook de waarde *n.a.* aannemen. Verder wordt de waarde 0 gebruikt om aan te geven dat er geen gelijkheid kan bestaan. Uit deze eigenschappen van de gelijkheidswaarde kan het gedrag van de operator worden afgeleid.

De operator berekent één gelijkheidswaarde uit meerdere gelijkheidswaarde. De gelijkheidswaarden tellen even zwaar: de operator berekent een gemiddelde over de gelijkheidswaarden die beschikbaar zijn. Door alleen een gemiddelde te berekenen over beschikbare waarden, is het bereik van alle vergelijkingen hetzelfde, namelijk in het interval [0,1].

Definitie. Gegeven *n* gelijkheidswaarden a_1, ..., a_n. Voor $a_1 \circ ... \circ a_n$ geldt:

$$a_1 \circ ... \circ a_n = \begin{cases} 0 & \exists a_i \in \{a_1,...,a_n\} : a_i = 0 \\ n.a. & \forall a_i \in \{a_1,...,a_n\} : a_i = n.a. \\ \dfrac{\sum\limits_{a_i \neq n.a.} a_i}{\sum\limits_{a_i \neq n.a.} 1}, \text{ met } i \in \{1,..,n\} & \text{anders} \end{cases}$$

Voorbeeld. Gegeven twee gelijkheidswaarden *a* en *b*. Voor $a \circ b$ geldt:
- als $a = 0 \vee b = 0$, dan $a \circ b = 0$
- als $a = n.a.$, dan $a \circ b = b$
- als $b = n.a.$, dan $a \circ b = a$
- in de overige gevallen geldt : $a \circ b = (a + b) / 2$

Stelling. De operator is commutatief ($a \circ b = b \circ a$) en niet associatief ($(a \circ b) \circ c \neq a \circ (b \circ c)$).

Bewijs. Voor de eerste twee regels in de commutativiteit eenvoudig na te gaan. Voor de laatste regel volgt dit uit de commutativiteit van de som-operator. Dat de operator niet associatief is, wordt bewezen door een tegenvoorbeeld. Stel, er zijn drie gelijkheidswaarden a=0.5, b=0.5 en c=1. Er geldt: $(a \circ b) \circ c = (1/2) \circ 1 = 0.5 \circ 1 = 1.5/2 = 0.75$ en $a \circ (b \circ c) = 0.5 \circ (1.5/2) = 0.5 \circ 0.75 = 1.25/2 = 0.625$. Dus $(a \circ b) \circ c \neq a \circ (b \circ c)$, de operator is niet associatief.

Gelijkheidswaarden uit kennisregels hebben een bepaald gewicht. In de huidige definitie telt elke waarde echter even zwaar. Stel een kennisregel heeft een gelijkheidswaarde *a* met een gewicht *g*. Dit wordt genoteerd als $a[g]$. De operator berekent nu een gewogen gemiddelde over de beschikbare gelijkheidswaarden. Er geldt:

Definitie. Gegeven *n* gewogen gelijkheidswaarden $a_1[g_1]$, ..., $a_n[g_n]$. Voor $a_1[g_1] \circ ... \circ a_n[g_n]$ geldt:

$$a_1[g_1] \circ ... \circ a_n[g_n] = \begin{cases} 0 & \exists a_i \in \{a_1,...,a_n\} : a_i = 0 \\ n.a. & \forall a_i \in \{a_1,...,a_n\} : a_i = n.a. \\ \dfrac{\sum\limits_{a_i \neq n.a.} g_i a_i}{\sum\limits_{a_i \neq n.a.} g_i}, \text{ met } i \in \{1,..,n\} & \text{anders} \end{cases}$$

Door invulling is eenvoudig na te gaan dat geldt: $a[1] \circ b[1] = a \circ b$.

3.3.1 Van attribuutgelijkheid naar entiteitgelijkheid

Kennisregels beschrijven de gelijkheid tussen twee gemeenschappelijke attributen. Bij de vergelijking van een entiteitcombinatie heeft een kennisregel één gelijkheidswaarde. Verder heeft elke kennisregel een gewicht (zie §3.1.3).

Gegeven een knoop K met *m* kennisregels. Elke kennisregel wordt genummerd: r_i en w_i leveren respectievelijk de gelijkheid en het gewicht van de i^e kennisregel.

Definitie. De entiteitgelijkheid van K wordt gedefinieerd als:

$$s_k = r_1[w_1] \circ ... \circ r_m[w_m]$$

De gelijkheidswaarde r_i kan berekend zijn uit meerdere gelijkheidswaarden, doordat het attribuut fysiek in een één-op-meer relatie met de entiteit is opgeslagen. Een voorbeeld hiervan is een persoonsattribuut dat onder delicten is opgeslagen. Door inconsistentie kan een enkelwaardig persoonsattribuut op die manier meerwaardig worden. Hierbij is het van belang om te bepalen welke attribuutwaarde het meeste voorkomt, aangezien deze waarde waarschijnlijk de juiste is. In veel gevallen van 'meerwaardige' attributen gedragen de waarden zich als entiteiten: elke waarde moet ook gevonden worden in de andere informatiebron. Het

berekenen van één gelijkheidswaarde voor meerwaardige attributen gebeurt daarom standaard ook met behulp van reconciliatie, zoals in paragraaf 3.4 wordt uitgewerkt. In andere gevallen kunnen andere methoden geschikter zijn. In Bijlage C worden enkele alternatieven besproken.

3.3.2 Van entiteitgelijkheid naar objectgelijkheid

De entiteitgelijkheid beschrijft de gelijkheid nog niet volledig. Uit de gelijkheidsdistributie is bekend dat alle gelijkheid bij de centrumknoop meetelt, maar dat de gelijkheid via andere knopen kunnen worden doorgegeven. Elke knoop kent daarom een vorm van objectgelijkheid. De objectgelijkheid bestaat uit de gelijkheid van:

- kennisregels van de eigen knoop (s_K);
- maximaal één ouderknoop ($S(0)$, bij geen ouderknoop: $S(0)$=*n.a.*);
- v virtuele relaties ($v \geq 0$, $S(i)$ met $i \in \{1,...,v\}$);
- n kindknopen ($n \geq 0$, $S(v+i)$ met $i \in \{1,...,n\}$).

Uit de gelijkheidswaarden wordt de objectgelijkheid berekend: $S(K) = S(0)\overbrace{\circ ... \circ S(v+n)}^{v+n>0} \circ s_K$. Een deel is optioneel en staat alleen in de formule als de conditie erboven geldt. De operator is commutatief, dus de volgorde in de formule maakt niet uit voor de uitkomst. Echter, omdat de eerste gelijkheidswaarde die 0 oplevert, de uitkomst vroegtijdig bepaalt (op 0), is het voordelig om de expressies waarvoor de minste rekenkracht nodig is vooraan te zetten.

Vanwege de hiërarchische structuur is de oudergelijkheid $S(0)$ bekend. Ook de gelijkheid uit virtuele relaties is al berekend, maar deze moet nog worden opgehaald. Zoals eerder besproken in §3.2.3, komen deze na S(0). De kindgelijkheid hoeft alleen berekend te worden als de overige gelijkheid groter dan 0 of onbepaald is. Deze gelijkheid komt daarom achteraan in de formule. De objectgelijkheid wordt nu berekend door:

$$S(K) = S(0)\overbrace{\circ S(1) \circ ... \circ S(v)}^{v>0} \circ s_K \overbrace{\circ S(v+1) \circ ... \circ S(v+n)}^{n>0}.$$

In deze formule bestaat de entiteitgelijkheid s_k uit de gelijkheidswaarden van de verschillende kennisregels. Stel een knoop K heeft m kennisregels. Verder geldt v=0 en n=1. Er zijn twee varianten om de gelijkheid van kennisregels in de formule op te nemen:

$$S(K) = S(0) \circ (r_1[w_1] \circ ... \circ r_m[w_m]) \circ S(1)$$
$$S(K) = S(0) \circ r_1[w_1] \circ ... \circ r_m[w_m] \circ S(1)$$

Aangezien de gelijkheidsoperator niet associatief is, kunnen beide varianten verschillende uitkomsten opleveren. De eerste variant berekent eerst de entiteitgelijkheid als een gelijkheidswaarde in het interval [0,1]. Daarna wordt de objectgelijkheid berekend. In de tweede variant telt elke kennisregel (gewogen) mee in de objectgelijkheid. Ten opzichte van de eerste variant zijn de verschillen:

- hoe meer kennisregels, hoe zwaarder weegt de entiteitgelijkheid
- hoe zwakker de kennisregels, hoe zwaarder wegen de andere objectgelijkheden

Dit komt overeen met de werkelijkheid: als een object zelf voldoende (sterke) eigenschappen heeft, dan zijn de overeenkomsten in gerelateerde objecten minder belangrijk. De tweede variant is daarom gekozen. De uiteindelijke definitie van objectgelijkheid luidt nu:

Definitie. Gegeven een knoop K met m kennisregels, n kindknopen en v virtuele relaties. De objectgelijkheid van K wordt gedefinieerd als:

$$S(K) = S(0)\overbrace{\circ S(1) \circ \ldots \circ S(v)}^{v>0}\overbrace{\circ r_1[w_1] \circ \ldots \circ r_m[w_m]}^{m>0}\overbrace{\circ S(v+1) \circ \ldots \circ S(v+n)}^{n>0}$$

In de formule voor objectgelijkheid moet verder gelden dat $m+n>0$, omdat oudergelijkheid alleen een object te weinig beschrijft.

De formule voor objectgelijkheid bestaat uit de entiteitgelijkheid en objectgelijkheid uit andere knopen. De objectgelijkheid uit andere knopen is één gelijkheidswaarde die de objectgelijkheid in de formule beschrijft. Zo beschrijft S(B) in S(A)=s_A∘S(B) de gelijkheid van entiteit-combinaties uit knoop B die een relatie hebben met de entiteitcombinatie uit knoop A waarvoor de objectgelijkheid berekend wordt. Als knoop B een kindknoop is van knoop A, dan kunnen dit meerdere entiteitcombinaties, en daarmee meerdere gelijkheidswaarden, zijn. Het berekenen van één gelijkheidswaarde uit meerdere gelijkheidswaarden wordt besproken in paragraaf 3.4.2.

3.4 Reconciliatie

Deze paragraaf behandelt eerst de reconciliatie van objectgelijke entiteiten. Daarna wordt besproken hoe met behulp van reconciliatie één gelijkheidswaarde berekend wordt uit meerdere gelijkheidswaarden.

3.4.1 Selecteren van de reconciliaties

Voor alle entiteitcombinaties is de objectgelijkheid berekend. De laatste stap is het conciliëren van gemeenschappelijke entiteittypen (de knopen). Hiervoor kunnen twee doelen worden opgesteld:

- het reconciliëren van objectgelijke entiteiten;
- het niet reconciliëren van de overige (objectongelijke) entiteiten.

Dey et al. ([DSD98], [DSD02]) maken ook gebruik van afstand bij het reconciliëren van entiteiten. Zij gebruiken een beslissingsmodel om de optimale conciliatie te bepalen, waardoor de conciliatie als een maximalisatieprobleem gezien wordt. De complexiteit van dit model is, bij respectievelijk m en n entiteiten in de eerste en tweede bron, in het ergste geval $O(N^3)$ met N=max{m, n}. In dit onderzoek hebben we echter te maken met grote aantallen vergelijkingen. Bovendien wordt er niet op één plaats gereconcilieerd, maar op verschillende plaatsen in het model (bij elke knoop, maar mogelijk ook bij attribuutwaarden). In dit onderzoek is een andere methode ontwikkeld, waarbij efficiëntie en kwaliteit gecombineerd worden.

Voor de reconciliatie worden alle combinaties van gelijksoortige entiteiten met elkaar vergeleken. De resultaten worden in een gelijkheidsmatrix geplaatst, waarin eerste dimensie (rijen) wordt gevormd door elementen uit bron I en de tweede dimensie (kolommen) door elementen uit bron II. Elke cel bevat de gelijkheidswaarde van de bijbehorende elementcombinatie.

Definitie. Een gelijkheidsmatrix *SM* is een 2-dimensionele matrix, waarin eerste dimensie (rijen) wordt gevormd door entiteiten uit bron I en de tweede dimensie (kolommen) door entiteiten uit bron II. Elke cel bevat de gelijkheidswaarde van de bijbehorende entiteitcombinatie. Bij een gelijkheidswaarde *n.a.* is de cel 0, omdat op basis van missende informatie een combinatie nooit gereconcilieerd kan worden.

Als een gelijkheidswaarde van een entiteitcombinatie boven de gelijkheidsdrempel uitkomt (gewoonlijk nul, maar dit kan theoretisch ook hoger zijn), dan is de entiteitcombinatie een reconciliatiemogelijkheid.

Definitie. Een reconciliatiemogelijkheid is een entiteitcombinatie met een gelijkheidswaarde groter dan de gelijkheidsdrempel. De gelijkheidswaarden *n.a.* en 0 leveren nooit een reconciliatiemogelijkheid op.

In de reconciliatiemethode wordt naast de gelijkheidswaarde ook de waarschijnlijkheid van een reconciliatiemogelijkheid gebruikt. De waarschijnlijkheid wordt gedefinieerd als kans. Binnen de kansrekening gelden de volgende regels:

1. de kans op een gebeurtenis is een waarde in het interval [0,1];
2. de kans op een gebeurtenis en zijn complementen tellen op tot 1;
3. als twee kansen A en B betrekking hebben op dezelfde gebeurtenis, dan sluiten de kansen elkaar uit: de kans dat beide kansen optreden is 0;
4. als twee kansen A en B betrekking hebben op verschillende gebeurtenissen, dan zijn de kansen onderling onafhankelijk (disjunct): $P(A \cap B)=P(A)P(B)$.

Gegeven een knoop K, een reconciliatiemogelijkheid (i, j) uit K en de m×n-gelijkheidsmatrix *SM* van K. Stel dat entiteit *i* met nog meer entiteiten een reconciliatiemogelijkheid heeft. De kans dat de reconciliatiemogelijkheid (i, j) voor entiteit *i* de juiste is, wordt mede bepaald door de gelijkheidswaarden van de andere reconciliatiemogelijkheden waar entiteit *i* deel van uitmaakt:

$$prob_i(i,j) = \frac{SM_{i,j}}{\sum_{k=1}^{m} SM_{k,j}}$$

Op dezelfde manier kan de kans $prob_j$ voor entiteit *j* berekend worden. Beide kansen voldoen aan de eerdergenoemde regels binnen de kansrekening.

De kansen voor entiteit *i* en *j* samen zijn disjunct, omdat ze in verschillende bronnen berekend worden. De voorwaardelijke kans van deze gebeurtenissen is daarom gelijk aan het product van de kansen.

Definitie. De waarschijnlijkheid of kans $prob(i, j)$ dat een reconciliatiemogelijkheid (i, j) de juiste is, wordt gedefinieerd als:

$$prob(i,j) = \frac{SM_{i,j}}{\sum_{k=1}^{m} SM_{k,j}} \cdot \frac{SM_{i,j}}{\sum_{k=1}^{n} SM_{i,k}}$$

De gelijkheidswaarde en waarschijnlijkheid worden gecombineerd in één waarde: de reconciliatiescore. Hierin heeft de gelijkheidswaarde een bepaald gewicht ten opzichte van de waarschijnlijkheid. Dit gewicht wordt aangegeven door de gelijkheidsfactor *sf*.

Definitie. De reconciliatiescore $rs(i, j)$, of kortweg score, van een reconciliatiemogelijkheid wordt gedefinieerd als:

$$rs(i,j) = \frac{SM_{i,j} \cdot sf + prob(i,j)}{sf + 1}$$

Meer onderzoek is nodig om een algemene uitspraak te doen over de juiste verhouding tussen gelijkheidswaarde en waarschijnlijkheid. In dit onderzoek is de gelijkheidsfactor geoptimaliseerd voor de beste resultaten. Uiteindelijk is een gelijkheidsfactor van 100 gebruikt.

De methode werkt als volgt: de reconciliatiemogelijkheid met de hoogste score gereconcilieerd, totdat de score onder een minimum is gekomen of totdat er geen mogelijkheden meer zijn. Als deze reconciliatiemethode wordt toegepast op meerwaardige attributen en de hoogste reconciliatiescore is niet uniek, dan wordt de reconciliatiemogelijkheid met de meeste voorkomens gereconcilieerd (voorkomens in bron I plus voorkomens in bron II).

Deze methode geeft geen garantie voor een conciliatie met minimale kosten zoals in het beslissingsmodel van Dey, maar komt daar naar verwachting wel in de buurt door steeds de mogelijkheid met de hoogste score te kiezen. De methode berekent de reconciliaties in één slag, waardoor efficiëntie maximaal is: de complexiteit van tijd is gelijk aan de complexiteit van grootte, namelijk $O(N^2)$.

3.4.2 Berekenen van één gelijkheidswaarde

In de vorige paragraaf zijn alle objectgelijke elementen gereconcilieerd. Voor de centrumknoop is dit het eindstadium. Op andere plaatsen moet deze informatie nog gebruikt worden om één gelijkheidswaarde te berekenen. Omdat het hier over zowel entiteiten als meerwaardige attributen kan gaan, wordt gesproken over elementen. Voor elke reconciliatiemogelijkheid is nu bekend of die gereconcilieerd is:

Definitie. Gegeven een reconciliatiemogelijkheid (i, j). De functie $recon(i, j)$ geeft de waarde 1 als de reconciliatiemogelijkheid gereconcilieerd is, en anders 0.

Gegeven een elementcombinatie (i, j). Laat $numocc_I(i)$ en $numocc_{II}(j)$ respectievelijk het aantal voorkomens van element i in bron I en van element j in bron II zijn. Laat verder $numocc_I$ en $numocc_{II}$ in de volgende definitie aflopende reeksen zijn.

Definitie. Het maximaal haalbare gewicht $w_{\max}$ is de grootste som van gewichten die gehaald kan worden bij het kiezen van reconciliatiemogelijkheden. Vanwege de reeksordening geldt voor een m×n-matrix:

$$w_{\max} = \sum_{i=1}^{\min(m,n)} (numocc_I(i) + numocc_{II}(i))$$

De som van de gewichten van gereconcilieerde elementen kan niet groter zijn dan $w_{\max}$. Een reconciliatiemogelijkheid wordt daarom gewaardeerd ten opzichte van $w_{\max}$.

Definitie. Gegeven een elementcombinatie (i, j). De waardering $occ(i, j)$ wordt gedefinieerd als:

$$occ(i, j) = \frac{numocc_I(i) + numocc_{II}(j)}{w_{\max}}$$

Definitie. Een samenvoegfunctie voegt de gelijkheid van een set van elementcombinaties K samen in één gelijkheidswaarde.

Definitie. De samenvoegfunctie voor gereconcilieerde gelijkheid $m(K)$ wordt gedefinieerd als:

$$m(K) = \sum (SM_{i,j} \cdot occ(i, j) \cdot recon(i, j))$$

Andere samenvoegfuncties met alternatieve berekeningen worden besproken in Bijlage C.

3.5 Conclusie

Om te bepalen of twee gelijksoortige entiteiten objectgelijk zijn, moet bepaald worden of de entiteiten naar hetzelfde object verwijzen. Bij gebrek aan een gemeenschappelijke sleutel kan dit alleen bepaald worden met behulp van de overlap tussen bronnen. De mate waarin eigenschappen overeenkomen wordt beschreven door middel van afstand. Hierdoor kan ook een mate van gelijkheid worden beschreven voor eigenschappen die semantisch ongelijk zijn, maar wel een bepaald verband hebben.

Schema-integratie en attribuutselectie is onmisbaar in het definiëren van de overlap. Door bronexperts de gemeenschappelijke eigenschappen te laten beschrijven op attribuutniveau, wordt de overlap betrouwbaar in kaart gebracht en wordt hiermee direct de schema-integratie en attribuutselectie afgehandeld.

Gemeenschappelijke eigenschappen beschrijven gelijkheid het beste wanneer ze geplaatst worden onder het bijbehorende gemeenschappelijke entiteittype (knoop). Dit hoeft niet de knoop te zijn waarvoor conciliatie gewenst is (de centrumknoop). Binnen elke knoop wordt entiteitgelijkheid berekend. De entiteitgelijkheid uit de knopen wordt gedistribueerd naar de centrumknoop en telt mee in de objectgelijkheid. De gebruiker bepaalt hiervoor een gelijkheidsdistributie, die feitelijk berust op de natuurlijke relaties die knopen hebben met de centrumknoop.

Door clustering wordt de efficiëntie van de berekening van entiteitgelijkheid verbeterd. De clustering gebeurt in een hiërarchische structuur; het informatiemodel. De gelijkheid in andere modellen wordt berekend met een gelijkheidsstelsel van meerdere informatiemodellen.

Expertkennis beschrijft gelijkheid op attribuutniveau. Deze gelijkheid wordt binnen een knoop omgezet in een entiteitgelijkheid. Alle entiteitgelijkheid in het informatiemodel bepaalt uiteindelijk de objectgelijkheid. Aan de hand van de berekende objectgelijkheid worden objectgelijke entiteiten gereconcilieerd. De ontwikkelde reconciliatiemethode reconcilieert entiteiten die het meest waarschijnlijk objectgelijk zijn. Dit gebeurt bovendien met minimale complexiteit ($O(N^2)$).

4 Casus

De ontwikkelde theorie wordt getoetst met behulp van een casus. De gebruikte casus bestaat uit drie informatiebronnen: HKS, OMDATA en OBJD. Deze bronnen zijn eerder besproken in paragraaf 2.1. De eerste paragraaf van dit hoofdstuk bespreekt hoe de expertkennis is verzameld en welke kennisregels de expertkennis heeft opgeleverd. Paragraaf 4.2 behandelt de modellering van gelijkheid in een informatiemodel. Tot slot wordt in de laatste paragraaf ingegaan op de gebruikte gegevens.

4.1 Beschrijving van gelijkheid

De expertkennis om gelijkheid te beschrijven is met standaardtechnieken ([Dur94], [SCG91], [Ste95]), waaronder interviews, van bronexperts verkregen. Een uitgebreide beschrijving hiervan is terug te vinden in de documentbijlage Onderzoeksomgeving. Dit heeft de volgende overlap opgeleverd:

Eigenschap	HKS	OMDATA
Geboortedatum	Geboortedatum	Geboortedatum
Geboorteland	Geboorteland	Geboorteland
Geslacht	Geslacht	Geslacht
Pleegdatum	Datum opmaak proces-verbaal	Pleegdatum
Wetsartikel	Eerste vijf wetsartikelen delict	Eerste vijf wetsartikelen aanklacht

Tabel 4.1 – Gevonden overlap

De gevonden overlap moeten geformaliseerd worden om de gelijkheid te kunnen beschrijven. Na de formalisatie kunnen de eigenschappen eenvoudig worden omgezet in kennisregels (zie §E.2.3, Toepassing kennisregels). Voor de duidelijkheid wordt hier al over de eigenschappen gesproken als kennisregels. In de formalisatie zijn de bronnen HKS en OMDATA respectievelijk genummerd met 1 en 2 ($attr_1$ en $attr_2$ betekent respectievelijk een attribuut uit HKS en een attribuut uit OMDATA).

Bij de formalisering van kennisregels moeten de volgende vragen beantwoord worden:

- Welke attributen worden vergeleken?
- Welke afstandsverdelingen zijn er?
- Voor elke afstandsverdeling:
    - Hoe wordt de afstand bepaald?
    - Welke formule beschrijft de verdeling?
    - Wat is het maximum van de formule?
    - Over welk subdomein van de attribuutwaarden gaat de afstandsverdeling? (alleen bij meerdere afstandsverdelingen)
    - Wat is de mate van inconsistentie of ruis?

Bij het formaliseren van een kennisregel is de plaats van het attribuut in het informatiemodel niet van belang.

Tijdens de analyse zijn bij sommige kennisregels nog enkele verbeteringen doorgevoerd. Deze verbeteringen worden ook direct in deze paragraaf besproken (onder *Optimalisatie*).

Geboortedatum

De kennisregel 'geboortedatum' vergelijkt twee geboortedatums. Als één van de datums onbekend is, dan kan er geen uitspraak gedaan worden. De gelijkheid wordt nu gedefinieerd door de gelijkheidsfunctie:

$$sim = \begin{cases} n.a. & \text{als } datum_1 = n.a. \vee datum_2 = n.a. \\ 0 & \text{als } datum_1 \neq datum_2 \\ 1 & \text{als } datum_1 = datum_2 \end{cases}$$

Optimalisatie

Geboortedatum is de meest discriminerende kennisregel, mits de geboortedatum bekend is. Het blijkt dat de snelheid van het reconciliatiescript sterk negatief beïnvloed wordt als de geboortedatum in één van de bronnen niet bekend is, omdat hiermee het aantal mogelijkheden sterk toeneemt. In HKS is deze negatieve invloed verzacht door er een extra attribuut bij te betrekken: geboortejaar. De referentieset (n=10.000) bevat 33 personen waarvan geen geboortedatum bekend is. Van deze personen is echter wel een geboortejaar bekend. De beschrijving van de geoptimaliseerde kennisregel wordt uitgebreid met een ruisdrempel over het geboortejaar:

$$sim = \begin{cases} n.a. & \text{als } datum_2 = n.a. \vee (datum_1 = n.a. \wedge jaar_1 = n.a.) \\ 0 & \text{als } datum_1 \neq datum_2 \\ 0.5 & \text{als } jaar_1 = jaar(datum_2) \\ 1 & \text{als } datum_1 = datum_2 \end{cases}$$

Geslacht

$$sim = \begin{cases} n.a. & \text{als } geslacht_1 = n.a. \vee geslacht_2 = n.a. \\ 0 & \text{als } geslacht_1 \neq geslacht_2 \\ 1 & \text{als } geslacht_1 = geslacht_2 \end{cases}$$

Geboorteland

$$sim = \begin{cases} n.a. & \text{als } gebland_1 = n.a. \vee gebland_2 = n.a. \\ 0 & \text{als } gebland_1 \neq gebland_2 \\ 1 & \text{als } gebland_1 = gebland_2 \end{cases}$$

Optimalisatie

$$sim = \begin{cases} n.a. & \text{als } gebland_1 = n.a. \vee gebland_2 = n.a. \\ sim_{land}(gebland_1, gebland_2) & \text{anders} \end{cases}$$

De geoptimaliseerde kennisregel is een virtuele relatie. De gelijkheidsfunctie sim_{land} verwijst naar het informatiemodel voor de reconciliatie van landen. Dit model wordt besproken in Bijlage D.

Pleegdatum

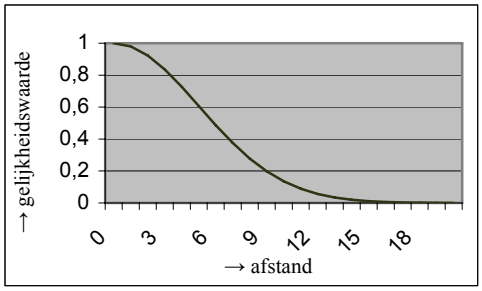
Deze kennisregel is het voorbeeld waarin gelijkheid wordt gedefinieerd door middel van een trendlijn. De datum waarop een delict gepleegd is ($datum_{delict}$, OMDATA) wordt vergeleken met de datum waarop het proces-verbaal is opgemaakt ($datum_{pv}$, HKS). De afstand tussen deze twee attributen wordt bepaald door de afstandsfunctie:

$$\delta = datum_{pv} - datum_{delict}$$

Een proces-verbaal kan niet eerder opgemaakt zijn dan de datum waarop een delict gepleegd is. Verder geldt dat de meeste processen-verbaal worden opgemaakt op de dag van een delict (d_0=0). De volgende gelijkheidsfunctie beschrijft de trendlijn die volgens de bronexperts het histogram van de verwachte afstanden beschrijft:

$$sim = \begin{cases} n.a. & \text{als } datum_{pv} = n.a. \vee datum_{delict} = n.a. \\ 0 & \text{als } \delta < 0 \\ e^{-\delta^2/50} & \text{als } \delta \geq 0 \end{cases}$$

De volgende figuur geeft de gelijkheidsfunctie grafisch weer.



Figuur 4.1 – Gelijkheidsfunctie kennisregel ‘pleegdatum’

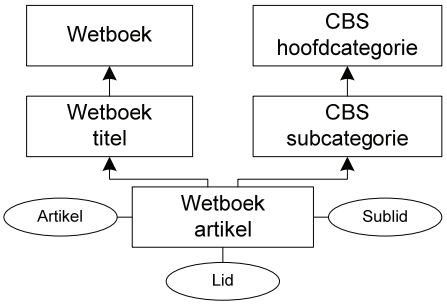
Het is bekend dat sommige delicten pas veel later bij de politie terechtkomen. Een grote afstand betekent dus niet automatisch ongelijkheid. Het is daarom nodig om een ruisdrempel in te bouwen. De gebruikte ruisdrempel ligt vlak boven de gelijkheidsdrempel (een verschil van 0,01), zodat de gelijkheid uit deze kennisregel altijd wordt meegenomen.

Optimalisatie

Slechts een klein deel van de delicten wordt laat bij de politie aangemeld. Het gaat hier waarschijnlijk ook om een bepaalde categorie zaken (zoals zwaardere vermogensdelicten). Dit is nog niet in detail onderzocht. Feit is wel dat bij zeer grote afstanden, de ruisdrempel op een bepaald moment niet meer genoeg waard is om ingezet te worden. Er zijn dan te weinig delicten die dankzij deze drempel nog gereconcilieerd worden. Bovendien wordt het aantal delicten dat ten onrechte bekeken wordt, te groot. Daarom er een maximumafstand ingesteld tot waar de ruisdrempel wordt ingezet. Deze afstand is gesteld op 1 jaar.

Wetsartikelen

Bij het vergelijken van wetsartikelen moet rekening worden gehouden met de mogelijkheid dat wetten herschreven worden en dat bovendien verschillende wetsartikelen over één delict kunnen gaan. Voor het vergelijken zijn de wetsartikelen daarom in een aantal categorieën ingedeeld. De indeling is schematisch weergegeven in Figuur 4.2.

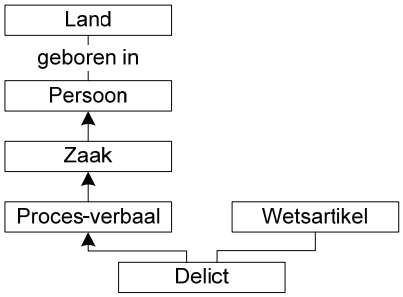


Figuur 4.2 – Schematische weergave indeling wetsartikel (in E/R diagram notatie)

Meer over de categorisering en reconciliatie van wetsartikelen is te vinden in Bijlage D.

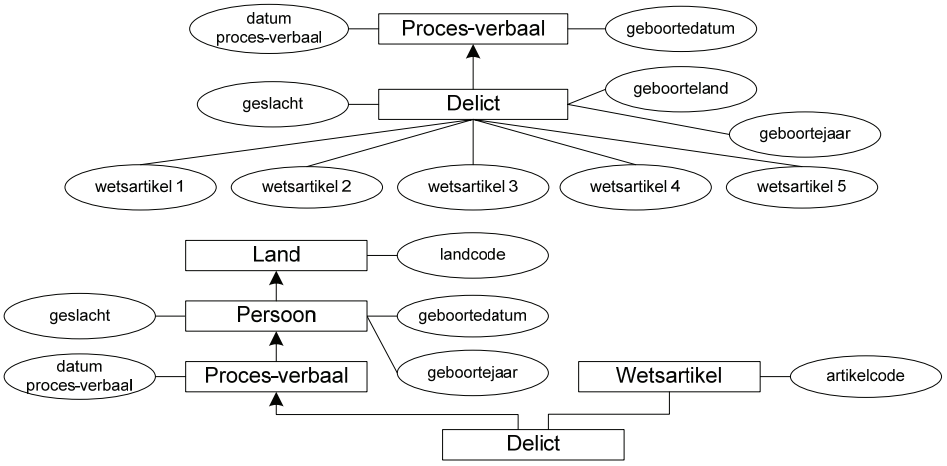
4.2 Modellering van gelijkheid

Uit de inventarisatie onder bronexperts en het analyseren van de informatiebronnen is het gemeenschappelijk datamodel afgeleid:

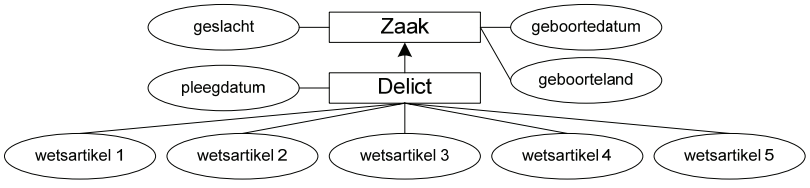


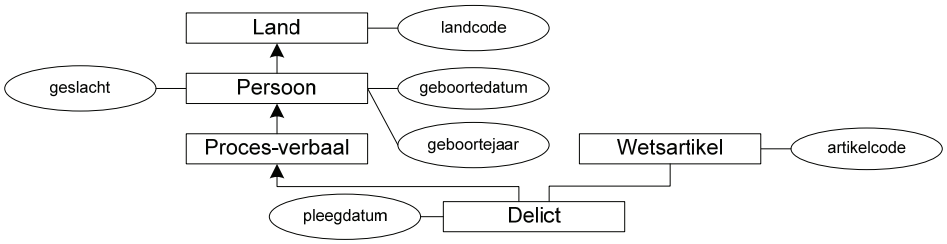
Figuur 4.3 – Gemeenschappelijk datamodel

Het datamodel wordt gevormd door de gemeenschappelijke entiteittypen waartoe de attributen van kennisregels behoren. Om de gemeenschappelijke entiteittypen te kunnen bepalen, is het van belang om de ligging van attributen in elke bron afzonderlijk te kennen, zowel fysiek als conceptueel (bij het bijbehorende entiteittype).



Figuur 4.4 – HKS fysiek (boven) en conceptueel (onder)





Figuur 4.5 – OMDATA fysiek (boven) en conceptueel (onder)

De datamodellen worden in de volgende paragraaf gebruikt om het informatiemodel vast te stellen.

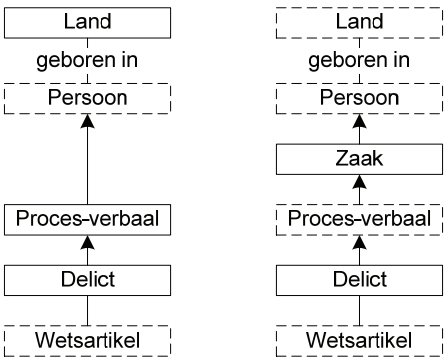
4.2.1 Informatiemodel

Om de conciliatie van twee bronnen te kunnen uitvoeren, moeten de attributen van elke kennisregel worden geplaatst onder een gemeenschappelijk entiteittype. In Figuur 4.4 en Figuur 4.5 is te zien dat dit alleen bij de kennisregel 'pleegdatum' niet het geval is. Ook sommige entiteittypen moeten nog worden afgeleid uit andere entiteittypen. Hierbij bevinden de attributen van het af te leiden entiteittype, inclusief de identificerende sterke sleutel, zich in een ander entiteittype. In figuren worden de afgeleide entiteittypen weergegeven met gestippelde blokken.

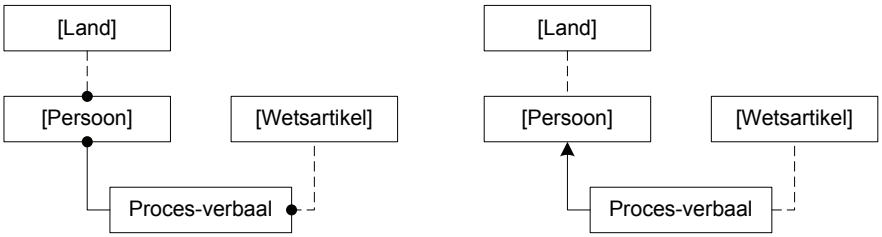
Bij de gemeenschappelijke entiteittypen Proces-verbaal en Delict speelt nog een ander probleem. Delicten vallen onder een proces-verbaal. Dit is een duidelijke één-op-meer relatie. De indeling kan echter nog wel eens verschillen. Het komt vaak voor dat twee delicten als twee processen-verbaal in HKS worden ingevoerd en later door het OM onder één proces-verbaal worden geplaatst. Een algemeen kenmerk van deze gevallen is vaak dat de delicten op dezelfde dag gepleegd zijn.

Ook bij wetsartikelen is geen sprake van een vast gegeven. Naarmate het strafproces vordert kunnen er wetsartikelen veranderen: soms komen er nieuwe bewijzen, waardoor een ander wetsartikel het delict beter beschrijft of de aanklacht versterkt kan worden. Reconciliatie van wetsartikelen onder een delict wordt dus ook uitgesloten.

Om bovenstaande redenen is besloten zowel de delicten als de wetsartikelen onder een delict op te laten gaan in Proces-verbaal. Merk op dat Wetsartikel als zelfstandig entiteittype wel te reconciliëren is.



Figuur 4.6 – Datamodellen van HKS en OMDATA met afgeleide entiteittypen



Figuur 4.7 – Gelijkheidsdistributie (l) en informatiemodellen (r)

Het behorende gelijkheidsstelsel *S* is:

$$S = \begin{cases} S(land) = s_{land} \\ S(wa) = s_{wa} \\ S(persoon) = (S(land)) \circ s_{persoon} \circ (s_{pv} \circ (S(wa))) \end{cases}$$

De scriptie richt zich op de conciliatie van het informatiemodel *S*(*persoon*). Hieraan voorafgaand moeten echter *S*(*land*) en *S*(*wa*) eerst geconcilieerd zijn. Dit wordt besproken in Bijlage D.

4.3 Gegevens

Dit onderzoek richt zich op het reconciliëren van personen in twee informatiebronnen: HKS en OMDATA. Van deze informatiebronnen is onbekend wat de correcte reconciliaties zijn. Om deze reden wordt er gebruik gemaakt van een referentieset, waarin de correcte reconciliaties wel bekend zijn. De eerste paragraaf beschrijft de referentieset, de tweede paragraaf het gebruik hiervan.

4.3.1 Beschrijving

De referentieset bestaat uit een representatieve 5%-steekproef van 10.000 persoonsentiteiten uit HKS uit het peiljaar 2000. Vervolgens zijn hierbij – voor zover mogelijk – persoonsentiteiten gezocht in OMDATA. Dit is voor 8.705 persoonsentiteiten gelukt. Bij de persoonsentiteiten in HKS zijn persoonsentiteiten in OMDATA gezocht. De referentieset bevat geen persoonsentiteiten in OMDATA die niet gereconcilieerd kunnen worden met een persoonsentiteit in HKS. In de praktijk kan dit wel voorkomen. De referentieset is vanuit HKS dus representatief, vanuit OMDATA niet. In de resultaatanalyse wordt daarom de nadruk gelegd op de resultaten gezien vanuit HKS.

De gemaakte reconciliaties in de referentieset zijn onafhankelijk. De reconciliaties blijven correct, ook na uitbreiding van de referentieset met andere persoonsentiteiten.

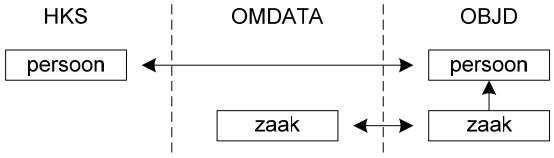
Voor de referentieset is extra persoonsinformatie gebruikt om de persoonsentiteiten handmatig te reconciliëren. Deze informatie, maar ook het proces zelf, kan daarom niet beschreven worden in dit document. De referentieset is met de grootst mogelijke nauwkeurigheid samengesteld. Daarom worden de correcte reconciliaties overgenomen als de waarheid.

4.3.2 Gebruik

De referentieset is op twee manieren in het onderzoek gebruikt:
- voor het creëren van een persoonsniveau in OMDATA;
- voor het toetsen van de resultaten.

Persoonsniveau in OMDATA

OMDATA bevat geen persoonsniveau, in tegenstelling tot de OBJD. Derhalve is de OBJD-informatie uit de referentieset gebruikt om het persoonsniveau te creëren in OMDATA. De referentieset bestaat uit de volgende koppelingen:

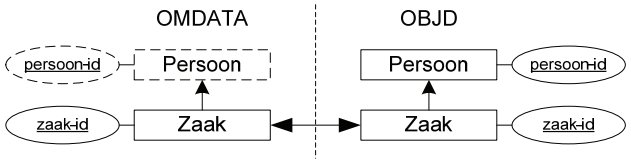


Figuur 4.8 – Koppelingen in de referentieset

Om het persoonsniveau vanuit OBJD in OMDATA te plaatsen, zijn uit de referentieset de OMDATA-zaken geleverd, uitgebreid met een uniek geanonimiseerd persoonsnummer (een soort volgnummer). Voor de toetsing zijn verder voor de gereconcilieerde persoonsentiteiten de bijbehorende metanummers uit HKS geleverd.

Om HKS en OMDATA op persoonsniveau te conciliëren, moeten personen in beide bronnen geïdentificeerd kunnen worden door middel van een sterke sleutel. In OMDATA is dit niet mogelijk. Via een zaakkoppeling met OBJD is daarom het persoonsniveau gecreëerd in OMDATA.

OMDATA en OBJD zijn beide afgeleid van COMPAS, waardoor het mogelijk is om zaken te koppelen middels een sterke sleutel. Via deze weg kan voor elke zaak in OMDATA een persoon worden vastgesteld. Schematisch:



Figuur 4.9 – Gemeenschappelijk datamodel van OMDATA en OBJD

Er zijn geen attributen gebruikt uit OBJD. Persoonskenmerken, zoals geboortedatum en geslacht, zijn dan ook alleen op zaakniveau beschikbaar. Hierdoor kan het voorkomen dat persoonskenmerken inconsistent zijn. Dit wordt in dit geval niet als nadeel bestempeld, omdat in HKS fysiek ook geen persoonsniveau beschikbaar is en de inconsistenties daar wellicht zijn terug te vinden (als ze uit een proces-verbaal zijn overgenomen). De inconsistenties zouden hierdoor zelfs kunnen leiden tot een betere score. In een later onderzoek kunnen de persoonskenmerken uit de OBJD wellicht als leidraad dienen om de juiste waarde te selecteren, maar dit valt buiten het onderzoekskader.

Toetsen van de resultaten

Daarnaast is de referentieset gebruikt om het resultaat van het reconciliatieproces te toetsen. Hiervoor zijn uit de referentieset de correcte reconciliaties overgenomen.

5 EROS

De centrale probleemstelling van dit onderzoek luidt: hoe kunnen entiteiten gereconcilieerd worden met beperkte overlap? In de voorgaande hoofdstukken is een theorie uiteengezet die dit mogelijk maakt. Deze theorie wordt getoetst aan de hand van een praktijkcasus. Hiervoor is het prototype EROS – Entity Reconciliation using Object Similarity – ontwikkeld.

5.1 Programma van eisen

Het prototype is een hulpmiddel om de doelstellingen van dit onderzoek te bereiken. Deze doelstellingen kunnen met het oog op de theorie opgedeeld worden in drie delen:

- het bepalen van het informatiemodel;
- het conciliëren van de centrumknoop door het berekenen van de reconciliatie-mogelijkheden en het maken van reconciliaties;
- het analyseren van de kwaliteit van de conciliatie.

De theorie behandelt dat het informatiemodel bestaat uit één of meer hiërarchische modellen, welke afzonderlijk (in een bepaalde volgorde) geconcilieerd moeten worden. EROS wordt gebruikt voor de conciliatie van één hiërarchisch model. Door EROS meerdere malen uit te voeren, kan het volledige informatiemodel geconcilieerd worden.

Bij het bepalen van het informatiemodel – en daarmee de gelijkheidsdistributie – moeten keuzes gemaakt worden in de richting waarin gelijkheid gedistribueerd wordt. Aangezien EROS slechts voor één casus wordt ontworpen, vormt dit proces geen onderdeel van EROS. In EROS geeft de gebruiker het hiërarchisch model (inclusief de centrumknoop) dat geconcilieerd moet worden aan.

Verder maakt het analyseren van de kwaliteit geen deel uit van het prototype. In het prototype zijn de precieze analyses nog niet bekend. Bovendien kunnen de analyses altijd achteraf handmatig gedaan worden. Er moet dan wel kwaliteitsinformatie worden opgeslagen. De doelstelling van het prototype luidt nu:

> "Het conciliëren van de centrumknoop door het berekenen van de reconciliatie-mogelijkheden en het maken van reconciliaties."

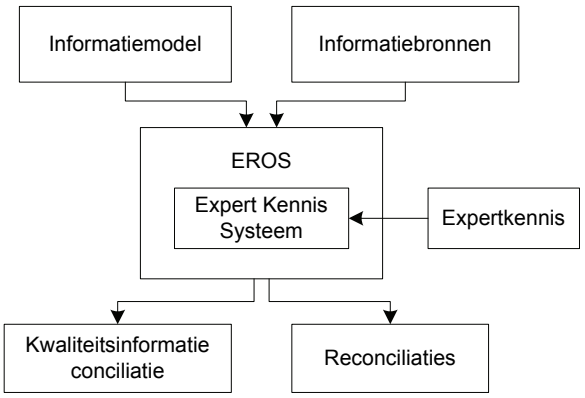
Om deze doelstelling te bereiken, moet EROS minimaal de volgende functionaliteit bevatten:

- Gebruikers moeten het hiërarchisch model, inclusief centrumknoop, kunnen aangeven.
- EROS zoekt reconciliatiemogelijkheden voor entiteitcombinaties van de centrumknoop.
- EROS maakt gebruik van expertkennis bij het zoeken naar reconciliatiemogelijkheden.
- EROS maakt een keuze in de reconciliatiemogelijkheden door de meest waarschijnlijke mogelijkheden te reconciliëren.
- EROS slaat kwaliteitsinformatie over de conciliatie op.

Vanwege het karakter van een prototype wordt in EROS alleen aandacht besteed aan gebruikersvriendelijkheid en snelheid, als dit de voortgang van het onderzoek bevordert.

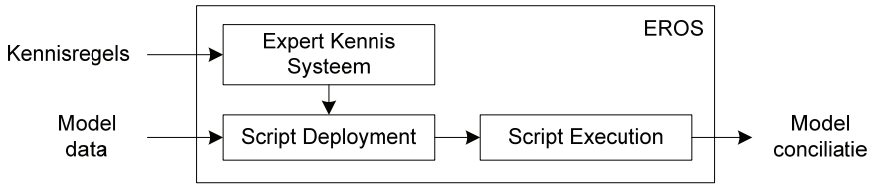
5.2 Architectuur

De gebruiker geeft het informatiemodel (inclusief centrumknoop) op. Verder gebruikt EROS de gegevens uit de informatiebronnen om gelijkheid te bepalen. De expertkennis die hierbij nodig is, wordt opgeslagen in een Expert Kennis Systeem. EROS produceert uiteindelijk reconciliaties, maar slaat ook andere reconciliatiemogelijkheden op, alsmede kwaliteits-informatie. De in- en uitvoer van EROS is schematisch weergegeven in Figuur 5.1.



Figuur 5.1 – EROS – invoer en uitvoer

De expertkennis wordt ingevoerd in een Expert Kennis Systeem. In dit prototype gebeurt dit niet door de experts zelf. Hierbij speelt namelijk de complexiteit van het geautomatiseerd formaliseren van expertkennis ([Dur94]). Dit valt buiten het onderzoekskader. De systeemarchitectuur van EROS is weergegeven in Figuur 5.2.



Figuur 5.2 – EROS – architectuur

De Script Deployment Tool genereert de benodigde scripts om de conciliatie uit te voeren. De scripts kunnen door de gebruiker worden uitgevoerd in de database. Tijdens het uitvoeren kan de voortgang worden gevolgd. Naderhand zijn de entiteiten gereconcilieerd en is er kwaliteitsinformatie opgeslagen over de kwaliteit van de conciliatie.

5.3 Ontwerp

5.3.1 Informatiemodel

Het informatiemodel bevat de gemeenschappelijke entiteittypen en de gebruikte relaties. Uit het informatiemodel en de centrumknoop kunnen de gelijkheidsdistributie en clustering worden afgeleid. Omdat EROS slechts een prototype is, wordt voor de invoering van het informatiemodel geen interface gebouwd. In plaats daarvan worden eisen gesteld aan de manier waarop de gegevens aangeleverd worden, zodat hieruit het informatiemodel kan worden afgeleid.

In het prototype wordt uitgegaan van entiteittypen met een identificerende sleutel van één numeriek attribuut. Door deze afbakening kan een entiteitcombinatie worden geïdentificeerd met twee attributen, wat de implementatie vereenvoudigt. Als een entiteittype niet aan deze eis voldoet, dan is dit eenvoudig te realiseren door elke entiteit een uniek nummer toe te kennen.

Om uit de gegevens het informatiemodel af te leiden, moeten de gegevens gestructureerd worden aangeboden. Door middel van SQL-views kunnen de gegevens getransformeerd worden in de gewenste structuur. Om de entiteiten te kunnen identificeren, heeft elk gemeenschappelijk entiteittype een eigen view in elke informatiebron. Een record in deze view representeert een entiteit. Meerwaardige attributen passen niet in deze view en hebben daarom een eigen view.

Definitie. Een *hoofdview* bevat de entiteiten van één entiteittype. Een meerwaardige attribuut staat in een *detailview*, die een één-op-meer relatie heeft met de hoofdview.

De views worden in de implementatie omgezet in tabellen, waarbij ook de nodige indices aangemaakt worden (*view materialisatie*). Als er geen transformatie nodig is, dan kan ook direct een tabel gebruikt worden. De gebruiker is dan zelf verantwoordelijk voor de benodigde indices.

Om de hiërarchische structuur te kunnen bepalen, moeten de hoofdviews de sleutel van de ouderknoop bevatten. De detailviews moeten een sleutel hebben die bestaat uit de sleutel van de bijbehorende knoop en het meerwaardige attribuut zelf. De sleutels moeten herkenbaar zijn voor het prototype. Voor de naamgeving van de sleutelattributen worden daarom de volgende richtlijnen opgesteld.

Definitie. Gegeven een entiteittype ET met een ouder EO en een meerwaardig attribuut EA. Er geldt:

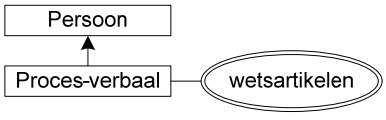
- In de hoofdview heeft de primaire sleutel de naam PK_ET.
- In de hoofdview heeft de vreemde sleutel naar de ouderknoop de naam FK_EO.
- In de detailview bestaat de primaire sleutel uit de attributen PK_ET en PK_EA.

De detailview kan verder nog een attribuut bevatten dat het aantal voorkomens van een attribuutwaarde bevat. Dit gegeven wordt gebruikt in de reconciliatie (zie §3.4). De structuur wordt verder toegepast in de naamgeving van de views:

- Een hoofdview wordt genoemd naar het entiteittype (ET).
- Een detailview wordt genoemd naar het entiteittype, gevolgd door een underscore (_) en een vrije tekst die de naam uniek maakt (ET_EA).

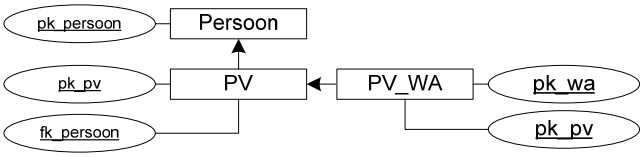
De hiërarchische structuur kan uit de sleutels worden afgeleid. De naamgeving van de views is slechts ter verduidelijking. In het hiërarchisch model kan een detailview de link zijn tussen twee entiteittypen met een meer-op-meer relatie en bevat de detailview de virtuele relatie naar het andere entiteittype. Hier wordt dieper op ingegaan in de implementatie (Bijlage E).

Het afleiden van het informatiemodel uit het fysieke model wordt gedemonstreerd aan de hand van een voorbeeld. Gegeven een hiërarchisch model met personen en processen-verbaal. Een proces-verbaal heeft één of meer wetsartikelen. Het gemeenschappelijke informatiemodel, waarin ook het conceptuele attribuut wetsartikelen is opgenomen, is:



Figuur 5.3 – Wetsartikel als meerwaardig attribuut (conceptueel)

Het fysieke datamodel (E/R diagramnotatie met sleutelattributen) wordt vervolgens:

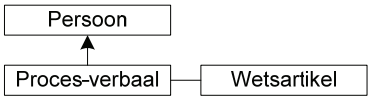


Figuur 5.4 – Wetsartikel als meerwaardig attribuut (fysiek)

Uit het fysieke datamodel kan het gemeenschappelijke informatiemodel worden afgeleid:

- Een hoofdview, horend bij een entiteittype, heeft één primair attribuut.
- De naam van het bijbehorende entiteittype kan afgeleid worden uit het primaire attribuut of de viewnaam.
- Een detailview heeft twee primaire attributen, waaronder het primaire attribuut uit de hoofdview.
- Persoon is een ouder van PV, vanwege de aanwezigheid van de primaire sleutel uit Persoon in de hoofdview PV.

Bovenstaand voorbeeld geeft wetsartikelen weer als meerwaardig attribuut. De wetsartikelen kunnen ook weergegeven worden als apart entiteittype met een meer-op-meer relatie met proces-verbaal:



Figuur 5.5 – Wetsartikel als entiteittype

Dit datamodel bestaat uit twee informatiemodellen, waarvan Wetsartikel als eerste geconcilieerd moet worden. In het informatiemodel Persoon bevat een detailview (PV_WA) de virtuele relatie naar Wetsartikel. Het fysieke model is precies hetzelfde als in het andere voorbeeld (zie Figuur 5.4).

De structuur geldt voor gemeenschappelijke entiteittypen en moet daarom voor beide informatiebronnen geïmplementeerd worden. Om duidelijk te maken uit welke bron de gegevens komen, hebben de views per bron een vaste prefix (bijvoorbeeld HKS_Persoon en OMDATA_Persoon).

5.3.2 Expert Kennis Systeem

Het Expert Kennis Systeem bevat kennisregels en de informatie om deze kennisregels toe te passen op informatiebronnen. Een kennisregel heeft de volgende eigenschappen:

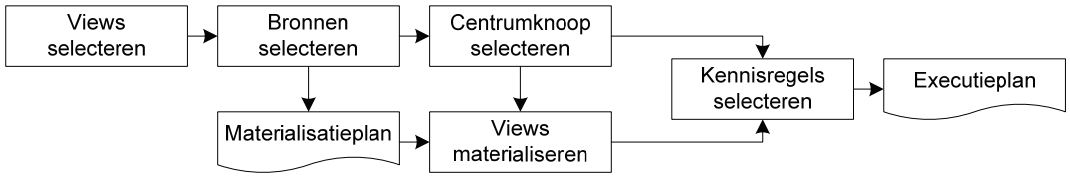
Eigenschap	Omschrijving
naam	De naam van de kennisregel
$bron_1$, $bron_2$	De bronnen waarvoor de kennisregel geldt
knoop	De knoop waar de kennisregel bij hoort
$attr_1$, $attr_2$	De attributen die de gemeenschappelijke eigenschap beschrijven
$aantal_1$, $aantal_2$	De attributen die het aantal voorkomens van de beschrijvende attributen bevatten
gelijkheidsfunctie	De functie die de gelijkheid tussen de attributen beschrijft.

Tabel 5.1 – Eigenschappen van een kennisregel

Verder kunnen er nog andere attributen een rol spelen, bijvoorbeeld bij het bepalen van de verschillende afstandsverdelingen.

5.3.3 Script Deployment

De eerste fase in het reconciliatieproces is het genereren van het executieplan, waarmee het reconciliatieproces wordt uitgevoerd. Deze fase bestaat uit de volgende stappen:



Figuur 5.6 – Stappen in Script Deployment

Allereerst selecteert de gebruiker de views (of tabellen) die het informatiemodel beschrijven. Uit de naamgeving worden vervolgens de informatiebronnen afgeleid. De gebruiker kiest twee informatiebronnen waartussen de reconciliatie plaats moet vinden.

Na het selecteren van de bronnen kan het informatiemodel afgeleid worden. De gelijkheidsdistributie kan worden afgeleid uit het informatiemodel na het selecteren van de centrumknoop. Nu is ook bekend welke views een rol kunnen gaan spelen in het reconciliatieproces. Het initiële gewicht van kennisregels hangt af van de statistieken van de bijbehorende attributen. Deze worden berekend door het materialisatieplan. Dit plan moet uitgevoerd worden voordat de kennisregels geselecteerd kunnen worden.

In het laatste stadium worden uit het Expert Kennis Systeem de kennisregels geselecteerd die gebruikt gaan worden in het reconciliatieproces. Alleen de kennisregels die gekoppeld zijn aan attributen van entiteittypen uit het gemeenschappelijk datamodel kunnen geselecteerd worden. Standaard worden alle geldige kennisregels geselecteerd met het initiële gewicht. De gebruiker heeft de mogelijkheid kennisregels te verwijderen uit de selectie of het gewicht handmatig in te stellen.

5.3.4 Script Execution

De tweede fase in het reconciliatieproces bestaat uit het uitvoeren van het materialisatie- en executieplan. EROS gebruikt hierbij drie soorten tabellen:

Soort	Prefix	Omschrijving
Materialisatietabellen	EROSM_	De materialisatie van views uit het informatiemodel (invoer)
Statistiektabellen	EROSS_	De statistieken van het reconciliatieproces: gelijkheidswaarden en gemaakte reconciliaties (uitvoer)
Statistiektabellen voor kennisregels	EROSR_	Statistieken van kennisregels: gelijkheidswaarden (uitvoer, handmatig)
Systeemtabellen	EROST_	Hulptabellen voor het berekenen van statistieken.

Tabel 5.2 – Tabelsoorten

Materialisatieplan

Het materialisatieplan materialiseert de views die het model beschrijven. Hierbij worden voor elke view de volgende stappen gevolgd:

1. Aanmaken materialisatietabel.
2. Aanmaken primaire sleutel en indices.
3. Kopiëren gegevens van view naar materialisatietabel.
4. Berekenen statistieken.

De laatste stap is nodig om de initiële gewichten van kennisregels uit te rekenen, welke nodig zijn in het executieplan.

Naast het materialiseren van views maakt het plan ook de EROS systeemtabellen aan; dit zijn tijdelijke hulptabellen voor het reconciliëren van gelijkheidsmatrices:

Systeemtabel	Inhoud
EROST_PROB	Bevat reconciliatiemogelijkheden en bijbehorende statistieken
EROST_PROB_SUM	Bevat de totale gelijkheid van alle reconciliatiemogelijkheden van een entiteit

Tabel 5.3 – Systeemtabellen

Executieplan

Het executieplan berekent voor elke entiteitcombinatie de reconciliatiemogelijkheden en kiest hieruit uiteindelijk maximaal één reconciliatiemogelijkheid als reconciliatie. Dit proces wordt het *reconciliatieproces* genoemd. Tijdens dit proces is er geen interactie met de gebruiker. Na afloop zijn de resultaten opgeslagen in statistiektabellen; één statistiektabel per knoop. De informatie die een statistiektabel kan bevatten is weergegeven in Tabel 5.4.

Attribuut	Betekenis
PK_I	Verwijzing naar de entiteit in bron I
PK_{II}	Verwijzing naar de entiteit in bron II
Similarity	De gelijkheid tussen de twee entiteiten
$Prob_I$	De waarschijnlijkheid van de mogelijkheid vanuit bron I
$Prob_{II}$	De waarschijnlijkheid van de mogelijkheid vanuit bron II
Reconciliated	Indicatie of de mogelijkheid gereconcilieerd is

Tabel 5.4 – Opbouw statistiektabel

Verder heeft de gebruiker de mogelijkheid om zelf statistieken op te slaan over de resultaten uit een specifieke kennisregel. Dit kan in statistiektabellen voor kennisregels, waarin de gelijkheid wordt opgeslagen die kennisregels opleveren. Elke knoop heeft een statistiektabel voor kennisregels.

Attribuut	Betekenis
PK_I	Verwijzing naar de entiteit in bron I
PK_{II}	Verwijzing naar de entiteit in bron II
Rule	Verwijzing naar de kennisregel
Similarity	De gelijkheid van de kennisregel

Tabel 5.5 – Opbouw statistiektabel voor kennisregels

5.4 Implementatie

De implementatie wordt besproken in Bijlage E. De bijlage behandelt onder andere de volgende onderwerpen:

- Implementatie van EROS
    - Modelaspecten (virtuele relaties, meerwaardige attributen)
    - Script Deployment (informatiemodel en kennisregels)
    - Script Execution (opbouw en mogelijkheden)
- Implementatie van de casus in EROS
    - Preparatie informatiemodel
    - Formalisatie kennisregels
    - Uitvoering van EROS
- Implementatie van de kennisregels vanuit het Kennis Expert Systeem (inclusief optimalisaties)

6 Resultaten

Om de casus in het prototype EROS te kunnen implementeren, zijn uitgebreide gesprekken gevoerd met bronexperts. Met de expertkennis die hieruit is voortgekomen, is het voor het eerst mogelijk geworden om de bronnen HKS en OMDATA te conciliëren op persoonsniveau. Er waren dan ook wisselende verwachtingen over de resultaten, al bestond het vermoeden dat een goede score zeker mogelijk moest zijn.

6.1 Inleiding

De behaalde resultaten zijn – gepresenteerd vanuit HKS en OMDATA – als volgt:

	HKS	OMDATA
Goed	93,8%	93,1%
Fout	6,2%	6,9%

Tabel 6.1 – Behaalde resultaten

De resultaten zijn behaald in drie ronden. In de eerste ronde is de expertkennis voor het eerst geïmplementeerd en zijn de bijbehorende kennisregels (geboortedatum, geslacht, geboorteland, antecedentdatum vs. pleegdatum en wetsartikelen) getest. In de tweede ronde zijn een aantal verbeteringen doorgevoerd. Er is een speciale samenvoegfunctie voor wetsartikelen geschreven (zie §E.2.3), de missende geboortedatums in HKS zijn ingevuld met geboortejaren en de ruisdrempel in de kennisregel "Antecedentdatum vs. Pleegdatum" wordt alleen ingezet als de afstand maximaal een jaar is. In de derde en laatste ronde is de bepaling van gelijkheid tussen landentiteiten verbeterd (zie Bijlage D). De verschillende ronden worden uitgebreid besproken in Bijlage F.

In dit hoofdstuk wordt de nadruk gelegd op de werkwijze die gevolgd kan worden bij het analyseren van de resultaten. Hierbij is vooral de laatste ronde interessant, omdat uit deze analyse direct aanbevelingen volgen voor mogelijkheden tot verbetering. Daarom worden in deze paragraaf alleen de resultaten van de laatste (derde) ronde besproken.

Intermezzo – Meerdere verbeteringen per ronde. Het reconciliatieproces is in verschillende ronden uitgevoerd. Van elke ronde zijn de resultaten geanalyseerd. Echter, in elke ronde zijn meerdere verbeteringen doorgevoerd. Is het niet beter om het resultaat van elke verbetering te analyseren?

Om deze vraag te beantwoorden, moet worden gekeken naar de reden van de verbetering. In dit onderzoek zijn twee redenen aan te wijzen die gebruikt zijn om een verbetering door te voeren. De belangrijkste is het beter beschrijven van de werkelijkheid. Daarnaast is één verbetering doorgevoerd om het aantal vergelijkingen te verminderen en hiermee de snelheid te verbeteren zonder de beschrijving van de werkelijkheid geweld aan te doen.

Bij een verbetering in de beschrijving van de werkelijkheid spelen de nieuwe resultaten geen rol in de beoordeling of de verbetering geslaagd is. De werkelijkheid wordt beter beschreven en als dit negatieve gevolgen heeft voor het resultaat, dan betekent dit zelfs dat eerdere vergelijkingen te optimistisch waren. Uiteraard moeten verbeteringen wel getest worden. Dit geldt met name voor verbeteringen in samenvoegfuncties, omdat hiervan het resultaat moeilijker te voorspellen is. Er geldt daarom dat er meerdere verbeteringen per ronde kunnen worden doorgevoerd, zolang er onafhankelijke testresultaten zijn. In ronde 2 is dit het geval, omdat elke verbetering zich in een andere knoop bevindt.

Bij een verbetering om het aantal vergelijkingen te verminderen wordt aangenomen dat de vergelijkingen de werkelijkheid nog steeds goed beschrijven. Of dit het geval is, moet in de nieuwe resultaten bekeken worden. Daarom geldt in dit geval ook dat de testresultaten onafhankelijk moeten zijn van andere verbeteringen om ze samen in één ronde te kunnen doorvoeren.

Het resultaat van de conciliatie wordt vergeleken met de reconciliaties in de referentieset. Hierbij worden entiteiten als uitgangspunt genomen: een entiteit kan namelijk wel of geen reconciliatie hebben in de referentieset. Daarom moeten de resultaten per bron bekeken worden.

Definitie. Een reconciliatie is correct als in de referentieset dezelfde reconciliatie gemaakt is.

Als er een reconciliatie gemaakt is en deze is correct, dan is het resultaat goed-positief. Als de gemaakte reconciliatie niet correct is, doordat er geen correcte reconciliatie bestaat of een andere reconciliatie correct is, dan is het resultaat fout-positief. Is er geen reconciliatie gemaakt en er is ook geen correcte reconciliatie, dan is het resultaat goed-negatief. Als geen correcte reconciliatie gemaakt is, terwijl deze wel bestaat, dan is het resultaat fout-negatief. Deze indeling kan worden weergegeven in een contingentietabel.

		Gelijk in werkelijkheid?	
		ja	**nee**
Gelijk in EROS?	**ja**	goed-positief	fout-positief
	nee	fout-negatief	goed-negatief

Tabel 6.2 – Contingentietabel voor resultaten

Bij de analyse is het interessant om te bekijken waarom een fout resultaat behaald is. Daarom worden de foutcategorieën ingedeeld naar de reden waarom de correcte reconciliatie (indien aanwezig) niet gekozen is:

reden fout	**fout-positief**	**fout-negatief**
niet aanwezig	onterecht gekozen	n.v.t.[6]
niet gekozen	verkeerd gekozen	niet gekozen
niet gevonden	niet gevonden (fout-positief)	niet gevonden (fout-negatief)

Tabel 6.3 – Indeling incorrecte resultaten

In de verdere bespreking wordt bovengenoemde indeling gebruikt. De twee varianten van de reden ‘niet gevonden’ zijn in de bespreking samengenomen in één categorie.

[6] Dit is een correct resultaat, namelijk goed-negatief.

6.2 Statistieken

Bij de bespreking van de resultaten worden de volgende definities gebruikt.

Definitie (herhaling). Een reconciliatiemogelijkheid, of kortweg mogelijkheid, is een entiteitcombinatie met een gelijkheidswaarde groter dan de gelijkheidsdrempel.
Definitie. Een reconciliatiealternatief, of kortweg alternatief, is een andere mogelijkheid dan de correcte mogelijkheid.

Na drie rondes zijn uiteindelijk de volgende resultaten behaald:

		HKS				**OMDATA**			
		Gelijk in werkelijkheid?				*Gelijk in werkelijkheid?*			
		ja		**nee**		**ja**		**nee**	
Gelijk in EROS?	**ja**	8108	81,1%	133	1,3%	8108	93,1%	133	1,5%
	nee	490	4,9%	1269	12,7%	464	5,3%	0	0,0%

Tabel 6.4 – Contingentietabel van de behaalde resultaten

Ingedeeld in goede en foute resultaten geeft dit:

Categorie	**HKS**		**OMDATA**	
Goed-positief	8108	81,1%	8108	93,1%
Goed-negatief	1269	12,7%	0	0,0%

Tabel 6.5 – Goede en foute resultaten

Categorie	**HKS**		**OMDATA**	
Niet gevonden (fout-negatief)	459	4,6%	440	5,1%
Niet gevonden (fout-positief)	11	0,1%	30	0,3%
Niet gekozen	31	0,3%	24	0,3%
Verkeerd gekozen	96	1,0%	103	1,2%
Onterecht gekozen	26	0,3%	0	0,0%

Tabel 6.6 – Foute resultaten

Voor persoonsentiteiten in OMDATA is er altijd een reconciliatiemogelijkheid. Dit is ook een gegeven uit de werkelijkheid: een persoon kan alleen in OMDATA voorkomen, als de persoon ook in HKS voorkomt. De indelingen ‘goed-negatief’ en ‘onterecht gekozen’ zijn voor OMDATA dan ook altijd leeg. Hierdoor is het aantal persoonsentiteiten in de categorie ‘goed-positief’ bovendien voor beide bronnen gelijk.

Een belangrijk percentage in het resultaat is het aantal correcte reconciliaties (goed-positief). Vanuit OMDATA is dit maximaal 100%, het proces behaalt 93,1%. Het foutpercentage van 6,9% kan worden uitgesplitst naar aantal reconciliatiemogelijkheden per entiteitcombinatie:

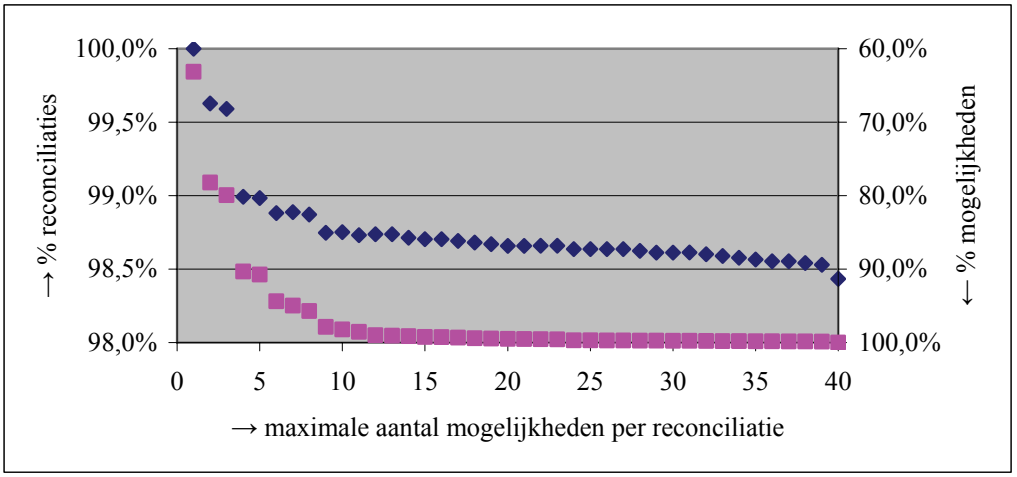
Aantal reconciliatiemogelijkheden	**Volume**	**Bijdrage aan foutpercentage**
0	5,3%	5,3%
> 3	20,7%	1,2%
2 of 3	14,3%	0,4%
1	59,7%	0,0%

Tabel 6.7 – Aantal reconciliatiemogelijkheden en bijdrage aan het foutpercentage

De belangrijkste oorzaak dat de correcte reconciliatie niet gevonden wordt, ligt in het feit dat de mogelijkheid niet gevonden wordt (5,3%). Zodra er een mogelijkheid gevonden wordt, wordt in 98,4% de juiste keuze gemaakt. Bij maximaal twee alternatieven ligt dit percentage zelfs op 99,6%. Uit deze statistieken kan, zonder diep in te gaan op de kwaliteit van de reconciliaties, al geconcludeerd worden dat het belangrijk is dat:

- reconciliatiemogelijkheden gevonden worden;
- het aantal reconciliatiealternatieven zo laag mogelijk wordt.

De volgende figuur ondersteunt de laatste conclusie:



Figuur 6.1 – Reconciliaties ten opzichte van de frequentie van het aantal mogelijkheden

De x-as toont het aantal reconciliatiemogelijkheden per reconciliatie. De linker y-as (en de bovenste grafiek) toont het percentage correcte reconciliaties bij maximaal *x* mogelijkheden. De rechter y-as (en de onderste grafiek) toont het percentage reconciliaties met maximaal *x* mogelijkheden ten opzichte van het totaal aantal reconciliaties.

Af te lezen is dat bij maximaal drie mogelijkheden per entiteitcombinatie (dit is 80% van het totaal) het percentage correcte reconciliaties 99,6% is. Het wordt echter ook duidelijk dat de frequentie van een bepaald aantal mogelijkheden afneemt naarmate het aantal groter wordt. Dat het percentage correcte reconciliaties blijft steken rond de 98% is vooral hieraan te danken. Een belangrijke conclusie is dat een laag aantal mogelijkheden een hoog percentage correcte reconciliaties tot gevolg heeft.

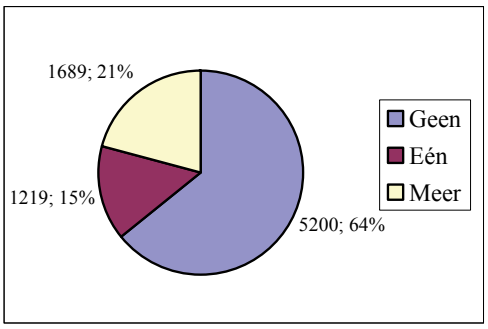
6.3 Kwaliteit

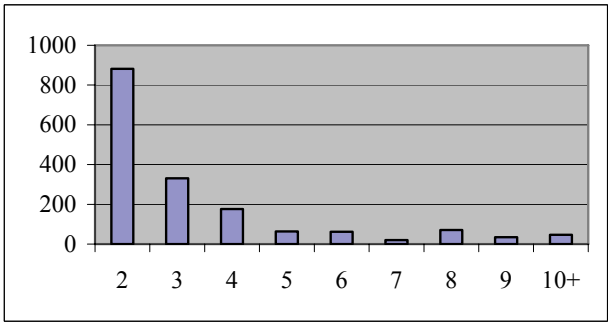
Om de kwaliteit van de conciliatie te beoordelen, moet gekeken worden naar de sterkte van de reconciliaties en de dreiging: het gevaar dat een verkeerde mogelijkheid gereconcilieerd wordt.

6.3.1 Goede resultaten

Goed-positief

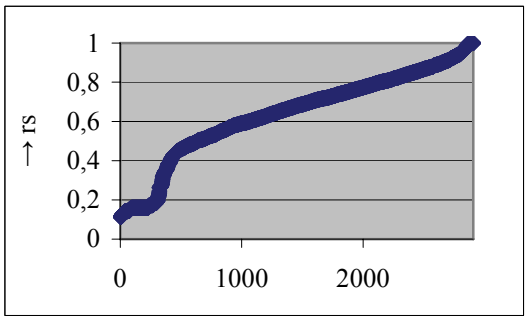
De waarschijnlijkheid dat een correcte reconciliatie gemaakt wordt, hangt af van het aantal alternatieven dat in het proces gevonden wordt. De grafiek in Figuur 6.2 geeft aan dat bij het merendeel van de reconciliaties (64%) geen alternatief gevonden wordt. Deze reconciliaties hebben de hoogst mogelijke waarschijnlijkheid om gekozen te worden.

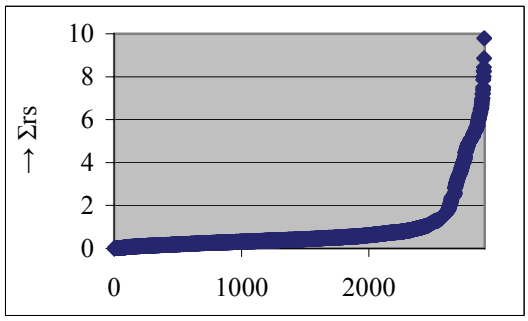




Figuur 6.2 (l) – Aantal alternatieven – 3 categorieën (absoluut en relatief)
Figuur 6.3 (r) – Aantal alternatieven – Categorie Meer (absoluut)

De overige reconciliaties hebben te maken met dreiging van alternatieven. Deze dreiging is het sterkst van het alternatief met de hoogste reconciliatiescore *rs* (Figuur 6.4). Maar ook de som van scores Σ*rs* speelt een rol, omdat hiermee de waarschijnlijkheden vanuit elke bron ($prob_{HKS}$ en $prob_{OMDATA}$) van de correcte reconciliatiemogelijkheid afnemen (Figuur 6.5).



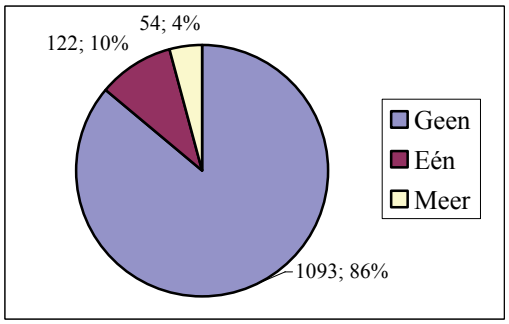


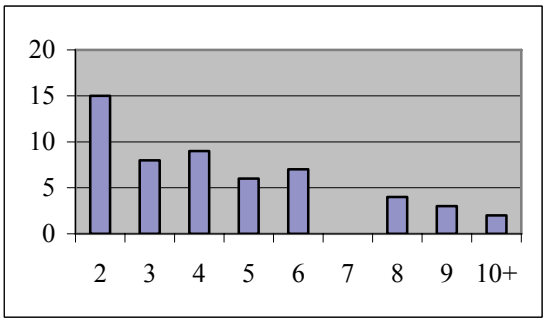
Figuur 6.4 (l) – Grootste dreiging onder de alternatieven (*rs*, 1 = maximale dreiging)
Figuur 6.5 (r) – Som van de dreiging van alle alternatieven (Σ*rs*, geen maximum)

Figuur 6.4 laat zien dat de dreiging van gevonden alternatieven sterk verschilt per correcte mogelijkheid. Hierover kunnen geen algemene uitspraken gedaan worden, behalve dat de dreiging kan worden verlaagd door te proberen de gelijkheid nog beter te beschrijven. Figuur 6.5 laat zien dat de dreiging van alle alternatieven laag is. Een figuur in deze vorm is wenselijk.

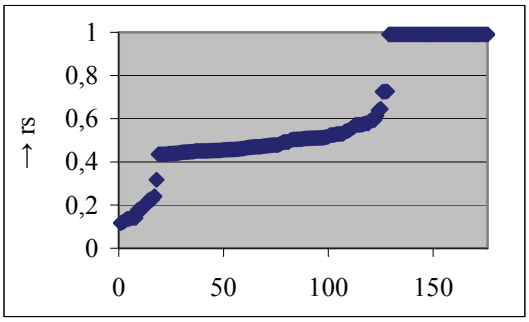
Goed-negatief

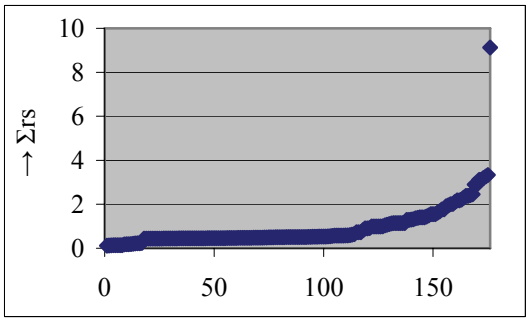
Voor deze categorie gelden dezelfde kwaliteitanalyses als voor de categorie ‘goed-positief’. Hier wordt alleen niet gesproken over dreiging van alternatieven, maar over dreiging van mogelijkheden, omdat in deze categorie geen correcte mogelijkheid bestaat.





Figuur 6.6 (l) – Aantal mogelijkheden – 3 categorieën (absoluut en relatief)
Figuur 6.7 (r) – Aantal mogelijkheden – Categorie Meer (absoluut)





Figuur 6.8 – Grootste dreiging onder de mogelijkheden (*rs*, 1 = maximale dreiging)
Figuur 6.9 – Som van de dreiging van alle mogelijkheden (Σ*rs*, geen maximum)

De dreiging van mogelijkheden is beperkt, zo blijkt uit Figuur 6.6 en Figuur 6.7. Een relatief groot aantal gevallen lijkt te maken te hebben met een mogelijkheid met maximale dreiging (Figuur 6.8). Uit nadere studie blijkt dat de dreiging net niet maximaal is. Dit verklaart waarom deze mogelijkheden niet onder 'fout gekozen' terecht zijn gekomen.

6.3.2 Foute resultaten

Niet gevonden

Als de correcte reconciliatiemogelijkheid niet gevonden wordt, dan hebben de objectgelijke persoonsentiteiten uit de twee bronnen niet voldoende gelijkheid. Deze entiteiten zullen nooit gereconcilieerd kunnen worden, zolang de mogelijkheid niet gevonden wordt. De analyse richt zich daarom vooral op de oorzaak van dit probleem. Het blijkt dat afwijkingen in de persoonskenmerken (geboortedatum, geslacht en geboorteland) een belangrijke oorzaak zijn.

Subcategorie	Aandeel
Afwijkingen in persoonskenmerken	46,6% [219 reconciliaties]
Overig	53,4% [251 reconciliaties]

Tabel 6.8 – Subcategorieën in categorie 'niet gevonden'

Afwijking	Percentage
Geboortedatum	54,3% [119 reconciliaties]
waarvan 1/1, 1/7	45,4% [54 reconciliaties]
Geboorteland	25,1% [55 reconciliaties]
Geslacht	25,1% [55 reconciliaties]

Tabel 6.9 – Afwijkingen in persoonskenmerken

Elk van de persoonskenmerken kan een afwijking vertonen, waardoor de som van deze percentages niet 100% is. Echter, in slechts 7 gevallen vertoont meer dan één kenmerk een afwijking.

In bijna de helft van de afwijkingen in geboortedatum is de datum respectievelijk 1 januari en 1 juli van hetzelfde jaar. Dit is een bekende inconsistentie in de politiegegevens: als alleen een geboortejaar van een verdachte bekend is, dan gebruiken sommige politiekorpsen 1 januari en andere 1 juli als geboortedag. In de overige gevallen is er meestal sprake van een mogelijke tikfout, bijvoorbeeld 20-3-1973 versus 29-3-1973.

Bij geslacht is de standaardwaarde tijdens de invoer waarschijnlijk 'mannelijk'. Vrouwen kunnen hierdoor per abuis als man geregistreerd worden.

In de derde ronde is de landvergelijking verbeterd. Deze verbetering zorgde bij 63 gevallen alsnog voor een reconciliatiemogelijkheid, zoals Tabel 6.10 laat zien.

Afwijking	Percentage
Geboortedatum	42,1% [119 reconciliaties]
waarvan 1/1, 1/7	45,0% [54 reconciliaties]
Geboorteland	41,1% [117 reconciliaties]
Geslacht	19,3% [55 reconciliaties]

Tabel 6.10 – Afwijkingen in persoonskenmerken in de 2e ronde

De overige afwijkingen hebben betrekking op de pleegdatum versus antecedentdatum en de wetsartikelen. Deze afwijkingen zijn bestudeerd door een willekeurige selectie van 10% beter te bekijken. In alle onderzochte gevallen bleken de verschillen in wetsartikelen genoeg om de reconciliatie als mogelijkheid af te keuren. De ruisdrempel van 1 jaar in de kennisregel 'pleegdatum versus antecedentdatum' speelde in deze selectie geen rol in de afkeuringen.

De persoonsentiteiten ondervinden dreiging van reconciliatiemogelijkheden die wel gevonden worden, waarbij deze persoonsentiteiten zelfs verkozen kunnen worden boven de correcte persoonsentiteiten. Dit is de groep 'niet gevonden (fout-positief)'.

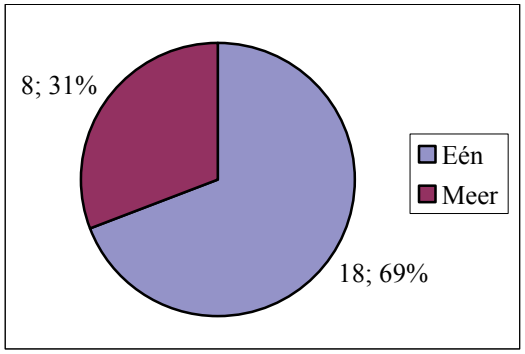
Resultaat dreiging	Aantal
Geen reconciliatie gemaakt	429
HKS-entiteit met verkeerde OMDATA-entiteit	11
OMDATA-entiteit met verkeerde HKS-entiteit	30

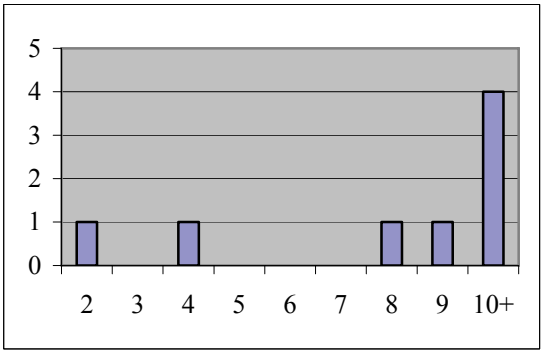
Tabel 6.11 – Resultaat van dreiging in de categorie 'niet gevonden'

Tabel 6.11 laat zien dat de aantallen in de groep 'niet gevonden (fout-positief)' beperkt zijn. In deze categorie is het bovendien belangrijker dat voor entiteiten eerst een reconciliatie-mogelijkheid wordt gevonden. Daarna kan opnieuw gekeken worden of de correcte reconciliatiemogelijkheid gekozen is.

Onterecht gekozen

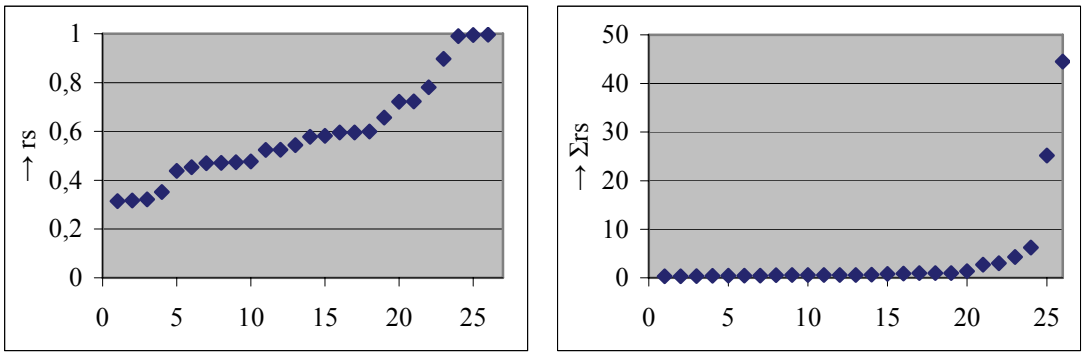
Bij persoonsentiteiten die gereconcilieerd worden terwijl er geen correcte reconciliatie-mogelijkheid is, spelen dezelfde factoren mee als in de categorie 'goed-negatief', met het verschil dat hier een reconciliatiemogelijkheid gekozen is. Daarom zijn ook dezelfde grafieken van toepassing.





Figuur 6.10 (l) – Aantal mogelijkheden –2 categorieën (absoluut en relatief)
Figuur 6.11 (r) – Aantal mogelijkheden – Categorie Meer (absoluut)

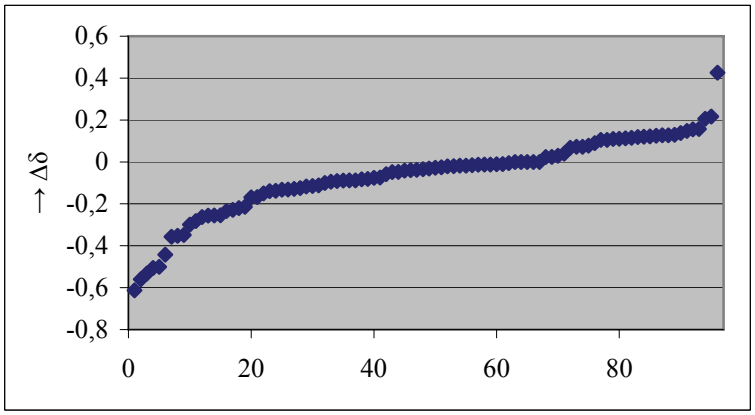
De dreiging bestaat uit de reconciliatiescore (*rs*) van de gekozen reconciliatie (Figuur 6.12) en de som van de score (Σ*rs*) van alle reconciliatiemogelijkheden (Figuur 6.13).



Figuur 6.12 (l) – Dreiging van de gekozen mogelijkheid (*rs*, 1 = maximale dreiging)
Figuur 6.13 (r) – Som van de dreiging van alle mogelijkheden (Σ*rs*, geen maximum)

Verkeerd gekozen

In deze categorie bevinden zich persoonsentiteiten die niet gereconcilieerd zijn met de correct persoonsentiteit, maar geconcilieerd zijn met een andere. Hier spelen twee reconciliatiemogelijkheden een rol en kan gekeken worden naar de afstand tussen deze twee mogelijkheden:

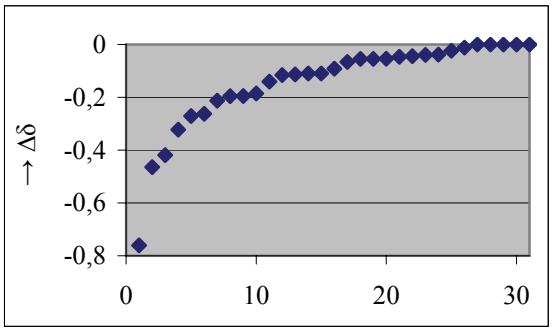


Figuur 6.14 – Afstand tussen correcte en gekozen mogelijkheid (correct minus gekozen; [-1,1])

Als de afstand groter dan nul is, dan heeft de gekozen mogelijkheid een kleinere reconciliatiescore dan de correcte mogelijkheid. Toch is de correcte mogelijkheid niet gekozen. Dit komt doordat de correcte mogelijkheid in andere gevallen een betere keus was. Als de afstand kleiner is dan nul, dan heeft de gekozen mogelijkheid een grotere reconciliatiescore dan de correcte mogelijkheid. Dit moet opgelost worden door het verbeteren van de implementatie van expertkennis. Als de afstand gelijk is aan nul, dan is de keuze puur willekeurig geweest. Ook dit kan alleen worden opgelost door het verbeteren van de implementatie van expertkennis.

Niet gekozen

Deze categorie bevat de persoonsentiteiten die – ondanks dat de correcte mogelijkheid gevonden is – niet gereconcilieerd zijn, ook niet met een andere mogelijkheid. Dit kan maar één oorzaak hebben: de andere helft van de reconciliatie is geconcilieerd met een andere mogelijkheid.



Figuur 6.15 – Afstand tussen correcte en gekozen mogelijkheid (correct minus gekozen; [-1,1])

Net als in categorie ‘verkeerd gekozen’ kan de afstand bepaald worden tussen de correcte en de gekozen mogelijkheid. Het blijkt dat voor alle mogelijkheden geldt dat de afstand kleiner dan nul is en de gekozen mogelijkheid dus beter is bevonden dan de correcte.

6.4 Conclusie

De behaalde resultaten zijn veelbelovend. Zodra een reconciliatiemogelijkheid gevonden wordt, wordt ruim 98% correct gereconcilieerd. Helaas is de informatie in de informatiebronnen niet in alle gevallen compleet of consistent, waardoor het uiteindelijke percentage dat correct gereconcilieerd wordt op 93% uitkomt.

De analyse heeft aangetoond dat een laag aantal reconciliatiealternatieven een positief effect heeft op het aantal correcte reconciliaties: bij maximaal twee reconciliatiealternatieven ligt het percentage correcte reconciliaties maar liefst op 99,6%. Het goede resultaat is in deze casus grotendeels toe te schrijven aan het lage aantal alternatieven: 90% van de gevallen heeft minder dan 5 alternatieven. Er is meer onderzoek nodig om de effecten op grotere informatiebronnen te voorspellen, omdat dan waarschijnlijk meer alternatieven een rol spelen. Op basis van de behaalde resultaten kan in elk geval geconcludeerd worden dat het belangrijk is dat:

- reconciliatiemogelijkheden gevonden worden;
- het aantal reconciliatiealternatieven zo laag mogelijk wordt.

Van ruim 5% waarvoor geen reconciliatiemogelijkheid gevonden wordt, kan dit worden toegeschreven aan twee oorzaken: missende informatie (delictgegevens) en inconsistente informatie (persoonsgegevens). Dit percentage kan alleen verminderd worden als de kwaliteit van de informatiebronnen zelf op deze punten verbeterd wordt.

De overige foute resultaten zijn klein in aantal (ruim 1%). Hiermee is de kwaliteit van de conciliatie voldoende voor dit onderzoek. Om dit percentage nog verder terug te dringen, moet de gelijkheid beter beschreven worden. Dit kan in eerste instantie door het verbeteren van de kennisregels. Als dit niet meer mogelijk is, dan moet gezocht worden naar mogelijkheden om nieuwe kennisregels toe te voegen (bijvoorbeeld door het binnenhalen van nieuwe attributen in de informatiebronnen).

De gepresenteerde grafieken geven inzicht in de kwaliteit van de conciliatie. De grafieken die betrekking hebben op dreiging door alternatieven of mogelijkheden hebben een gewenste vorm, waarbij de dreiging zo klein mogelijk blijft. Door specifieke probleemgevallen te bekijken die in de grafieken voor een ongewenste vorm zorgen, kunnen problemen met de kwaliteit gedetecteerd worden. Het oplossen van probleemgevallen leidt een kwaliteitsverbetering van de conciliatie in het algemeen.

7 Conclusies en aanbevelingen

In dit hoofdstuk wordt geprobeerd een antwoord te geven op de onderzoeksvragen die ten grondslag liggen aan dit onderzoek. De centrale probleemstelling luidt:

“Hoe kunnen objectgelijke entiteiten gereconcilieerd worden ondanks beperkte overlap?”

De centrale probleemstelling is opgedeeld in drie onderzoeksvragen:

- Hoe kan de overlap tussen twee informatiebronnen gedefinieerd worden?
- Hoe kan de overlap gebruikt worden in het reconciliëren van objectgelijke entiteiten?

- Hoe kan de kwaliteit van de conciliatie en de gebruikte gegevens uit informatiebronnen bepaald worden?

In de volgende paragraaf worden eerst de onderzoeksvragen en tot slot de centrale probleemstelling besproken. Paragraaf 7.2 geeft aanbevelingen op het gebied van toekomstig werk en de toepassing van het huidige onderzoek.

7.1 Conclusies

De overlap tussen twee informatiebronnen wordt gedefinieerd door bronexperts gemeenschappelijke eigenschappen op attribuutniveau te laten beschrijven. Door eigenschappen te beschrijven door middel van afstand kunnen eigenschappen die semantisch ongelijk zijn, maar wel een bepaald verband hebben, ook als overlap gedefinieerd worden.

Door gemeenschappelijke eigenschappen te plaatsen onder het bijbehorende gemeenschappelijke entiteittype (knoop) wordt gelijkheid het beste beschreven. De gelijkheid in elke knoop wordt gedistribueerd naar het centrale gemeenschappelijke entiteittype, waarvoor conciliatie gewenst is: de centrumknoop. De relaties tussen knopen in de gelijkheidsdistributie worden beschreven in een informatiemodel. De gelijkheid in een knoop kan efficiënt berekend worden door clustering in een hiërarchisch informatiemodel. De gelijkheid in andere modellen wordt berekend met een gelijkheidsstelsel van meerdere informatiemodellen.

De berekende gelijkheid komt samen in de centrumknoop en vormt zo de objectgelijkheid tussen twee entiteiten. De entiteitcombinatie die het meest waarschijnlijk objectgelijk zijn worden gereconcilieerd met een efficiënte methode. In de empirische fase is deze methode ook effectief gebleken; ruim 98 procent wordt correct gereconcilieerd als de correcte entiteitcombinatie als reconciliatiemogelijkheid gevonden wordt.

De kwaliteit van de conciliatie en de gebruikte gegevens wordt afgeleid uit de statistieken die zijn opgeslagen tijdens het reconciliatieproces. Correcte reconciliatiemogelijkheden die niet gevonden worden leiden naar inconsistente of missende gegevens; of naar verbeteringen in de beschrijving van de overlap. De kwaliteit van de conciliatie wordt afgemeten aan het foutpercentage, maar ook aan de dreiging van foute reconciliatiemogelijkheden.

Geconcludeerd kan worden dat entiteitreconciliatie ondanks beperkte overlap goed mogelijk is door middel van objectgelijkheid. Door de gelijkheidsdistributie kunnen alle gemeenschappelijke eigenschappen die tussen twee informatiebronnen bestaan effectief gebruikt worden om te bepalen of twee gelijksoortige entiteiten objectgelijk zijn. Mede door gebruik van clustering in de berekening van gelijkheid worden reconciliaties efficiënt berekend, ondanks de beperkte overlap van slechts vijf gemeenschappelijke eigenschappen.

7.2 Aanbevelingen

Op basis van de conclusies uit de vorige paragraaf en de opgedane ervaringen tijdens het onderzoek worden enkele aanbevelingen en suggesties voor toekomstig werk gedaan.

- Gebruiken van expertkennis

  Er moet voldoende kennis in het systeem aanwezig zijn om de reconciliaties betrouwbaar genoeg te maken. Uit de referentieset blijkt dat met expertkennis, mits de gegevens consistent zijn, stabiele en hoge scores worden gehaald: ongeacht het aantal reconciliatiemogelijkheden is het percentage correcte reconciliaties ruim 98%. Pas wanneer het aantal reconciliatiemogelijkheden minder dan 4 is, maakt het percentage een sprong naar boven de 99,5%. Aanbevolen wordt om zoveel expertkennis te verzamelen om het aantal reconciliatiemogelijkheden zo laag mogelijk te houden.

  Vanwege de korte looptijd van dit onderzoek is de expertkennis in de casus deels gebaseerd op experts, maar ook deels op de inhoud van de referentieset en/of de informatiebronnen zelf. Met meer onderzoek kan de expertkennis verder worden aangescherpt om de werkelijkheid beter te beschrijven. Ook is niet uitgesloten dat er nog meer expertkennis te vinden is. Voor het WODC is een belangrijke aanbeveling om verder te gaan met het in kaart brengen van expertkennis.

- Oplossen van inconsistenties

  In het algemeen kan worden opgemerkt dat inconsistenties een grote invloed kunnen hebben op de correctheid van de conciliatie. Als van een attribuut zowel de juiste als een onjuiste waarde bekend is, dan kan dit door het reconciliatieproces worden opgevangen. Als van een attribuut echter alleen een onjuiste waarde bekend is, dan kan de correcte reconciliatiemogelijkheid nooit worden gevonden. In de referentieset is bij 5,3% uiteindelijk geen enkele reconciliatiemogelijkheid gevonden, terwijl slechts in 1,6% een foute keuze in de reconciliatiemogelijkheden wordt gemaakt. Inconsistenties spelen een grotere rol naarmate ze dichter bij de wortel van het informatiemodel zitten, omdat ze dan invloed hebben op meer entiteiten. Zij kunnen daarom ook het beste vanaf de wortel worden aangepakt.

- Zoeken naar nieuwe overlap (data mining)

  EROS is in staat om statistieken bij te houden van kennisregels. In dit onderzoek komen deze kennisregels uitsluitend uit het Kennis Expert Systeem. Een mogelijke uitbreiding zou kunnen zijn dat EROS zelf op zoek gaat naar betrouwbare overlap tussen attributen en de uitkomsten rapporteert aan experts. Dit kan alleen als met grote zekerheid bekend is welke reconciliaties correct zijn, zoals bijvoorbeeld bij de referentieset in dit onderzoek het geval

is. De gevonden resultaten kunnen vervolgens door experts worden gebruikt om nieuwe kennisregels te ontwikkelen, of worden gebruikt in verdere prognoses. In deze nieuwe toepassing van de theorie is het voorspellen van het extrapolatiegedrag nog belangrijker dan in de huidige toepassing van de theorie. Er wordt in deze toepassing kennis opgedaan die niet meer berust op de werkelijkheid, maar op het voorkomen in de referentieset. Het is daarom nog moeilijker om te bewijzen dat deze kennis ook in een andere (grotere) set geldt. Daarnaast is het in deze toepassing noodzakelijk om bij verdere prognoses de ingevoerde expertkennis te kennen. Expertkennis uit door EROS gevonden overlap mag geen conclusie in prognoses worden.

- Zoeken naar minimale overlap (privacy)

  Eén van de aanleidingen voor dit onderzoek was de behoefte om personen te reconciliëren bij gebrek aan gemeenschappelijke sleutels of persoonskenmerken. Dit onderzoek heeft aangetoond dat, zeker voor een kleine informatieset, zeer weinig overlap nodig is om betrouwbare reconciliaties te maken. EROS kan gebruikt worden als applicatie om de privacygevoeligheid van informatiebronnen te bepalen. Zodoende is het mogelijk om een keuze te maken of de informatie anoniem genoeg is om te verstrekken aan een externe partij. Toekomstig onderzoek kan zich richten op het bepalen van de minimale overlap die nodig is voor betrouwbare persoonsreconciliatie.

- Toepassen buiten een referentieset (extrapolatie)

  Met de gebruikte casus kon niet onderzocht worden hoe het extrapolatiegedrag van het algoritme vanuit een referentieset voorspeld kan worden. Verwacht wordt dat, bij gebruik van EROS voor reconciliatie op basis van objectgelijkheid, dit gedrag tot op zekere hoogte voorspeld kan worden. Toekomstig onderzoek kan zich richten op de mogelijkheden om het gedrag te voorspellen en de betrouwbaarheid van de voorspelling aan te geven. In Bijlage G wordt hier een aanzet voor gegeven.

- Gebruik van betrouwbaarheidsintervallen

  Het beschrijven van de gelijkheid gebeurt onder andere door middel van afstandsverdelingen. In het onderzoek wordt gesteld dat het wenselijk is dat de verwachte afstand zich rond één afstand concentreert. Hoe meer afstanden bij de verwachte afstand in de buurt komen (en hierdoor de frequenties van die afstanden in de buurt komen van de maximale frequentie van de verwachte afstand), hoe sterker de vergelijking. Een betrouwbaarheidsinterval rond de verwachte afstand kan voorkomen dat te zwakke vergelijkingen worden ingevoerd. In dit onderzoek speelde dit probleem niet, maar toekomstig onderzoek kan hier verder op in gaan.

- Verder ontwikkelen van het prototype

  EROS is in dit onderzoek gebouwd als prototype waarin de theorie volledig toegepast kan worden, maar vereist verdere ontwikkeling voordat het ingezet kan worden in een gebruikersomgeving. Hierbij valt te denken aan een semi-geautomatiseerd Expert Kennis Systeem, ondersteuning bij het opstellen van de gelijkheidsdistributie, ondersteuning bij het opstellen en vullen van de informatiemodellen en ondersteuning bij het analyseren van de resultaten.

- Verbeteren van de theorie

  De theorie die ontwikkeld is in dit onderzoek werkt met twee verschillende informatiebronnen. Het ondersteunen van meer dan twee informatiebronnen zou een waardevolle uitbreiding zijn op de theorie. Verder worden in het reconciliatieproces een aantal parameters en methodes gebruikt. De huidige instellingen zijn geoptimaliseerd voor de gebruikte referentieset. Toekomstig onderzoek kan zich richten op een betere beschrijving van de invloed die deze instellingen hebben op het resultaat van het reconciliatieproces.

- Verbeteren van de efficiëntie

  Het onderzoek is uitgevoerd met een relatief kleine referentieset. Om het resultaat van het onderzoek ook in te zetten voor grotere sets is meer onderzoek nodig. Zo is het reconciliatieproces nog te traag om te kunnen werken met grote aantallen. Hiervan kan een groot deel worden toegeschreven aan de beperkingen van de gebruikte taal (PL/SQL), dus het uitbesteden van een deel van de code naar een externe applicatie is mogelijk al een oplossing.

Bij aanvang van dit onderzoek was het WODC geïnteresseerd in de mogelijkheden om informatiebronnen ondanks beperkte overlap te koppelen. Binnen het kader van een referentieset heeft dit onderzoek daar een antwoord op gegeven in de vorm van reconciliatie door middel van objectgelijkheid; onderzoek naar toepassing buiten dit kader (extrapolatie) is voor het WODC de logische volgende stap. Daarnaast blijkt uit de resultaten dat de kwaliteit van de informatiebronnen nog verbeterd kan worden. Helaas is het oplossen van de kwaliteitsproblemen waarschijnlijk geen optie, omdat de informatiebronnen onderhouden worden door externe partijen. Het verder in kaart brengen van de expertkennis zou een oplossing kunnen leveren in de vorm van nieuwe, betrouwbaardere kennisregels.

In het algemeen biedt de theorie in dit onderzoek een interessante basis voor meer onderzoek in de richting van data mining en privacy-gerelateerde toepassingen. Het prototype EROS is geschikt voor het toepassen van de theorie, maar vereist nog meer aandacht op het gebied van gebruikersvriendelijkheid en efficiëntie. Een terrein dat ook voor het WODC interessant is, is het uitbreiden van de theorie (en EROS) naar ondersteuning van een conciliatie met meer dan twee informatiebronnen.

Referenties

[AKWS95] S. Agarwal, A.M. Keller, G. Wiederhold en K. Saraswat, "*Flexible Relation: An approach for integrating data from multiple, possibly inconsistent databases*", IEEE, 1995.

[BDS03] P. Bonatti, Y. Deng, V.S. Subrahmanian, "*An Ontology-Extended Relational Algebra*", IEEE, 2003.

[BOT86] Y. Breitbart, P.L. Olson en G.R. Thompson, "*Database integration in a distributed heterogeneous database system*", Proc. Int. Conf. Data Eng., pp. 301-310, 1986.

[CTK96] A.L.P. Chen, P.S.M. Tsai en J.-L. Koh, "*Identifying Object Isomerism in Multidatabase Systems*", Distributed and Parallel Databases, vol. 4, nr. 2, pp. 143-168, april 1996.

[CBF04] R.S. Choenni, H.E. Blok en M.M. Fokkinga, *"Extending the Relational Model with Uncertainty and Ignorance"*, Technical report, juli 2004, no. TR-CTIT-04-29, pp. 1-16, Centre for Telematics and Information Technology (CTIT), Universiteit Twente, Nederland, ISSN 1381-3625

[CBL06] R.S. Choenni, H.E. Blok en E.C. Leertouwer, *"Handling Uncertainty and Ignorance in Databases: A Rule to Combine Dependent Data"*, Proc. 11th Int. Conf. on Database Systems for Advanced Applications (DASFAA), april 2006.

[Coh98] W.W. Cohen, "*Integration of Heterogeneous Databases without Common Domains using Queries Based on Textual Similarity*", Proc. 1998 ACM SIGMOD Conf., pp. 201-212, juni 1998.

[CER87] B. Czejdo, D.W. Embley en M Rusinkiewicz, "*An approach to schema integration and query formulation in federated database systems*", Proc. 3rd Int. Conf. Data Eng., pp. 477-484, 1987.

[DeM89] L.G. DeMichiel, "*Resolving database incompatibility: An approach to performing relational operations over mismatched domains*", IEEE, 1989.

[DSD98] D. Dey, S. Sarkar en P. De, "*A Probablistic Decision Model for Entity Matching in Heterogeneous Databases*", Management Science, vol. 44, nr. 10, pp. 1379-1395, 1998.

[DSD02] D. Dey, S. Sarkar en P. De, "*A Distance-Based Approach to Entity Reconciliation in Heterogeneous Databases*", IEEE Trans. Knowledge and Data Eng., vol. 14, no. 3, pp. 567-582, 2002.

[Dur94] J. Durkin, "*Expert Systems – Design and Development*", Prentice Hall International, 1994.

[EH04] A.Th.J. Eggen en W. van der Heide, "*Criminaliteit en rechtshandhaving 2004*", Boom Juridische uitgevers, 2004, ISBN 90-5454-642-5.

[Gass85] S. Gass, "*A Process to Determine Priorities and Weights for Large-Scale Linear Goal Programming*", Proc. 12th Int. Symp. Math. Programming, augustus 1985.

[HS95] M.A. Hernández en S.J. Stolfo, "*The Merge/Purge Problem for Large Databases*", Proc. 1995 ACM SIGMOD Conf., pp. 127-138, mei 1995.

[HSA03] T. Hussain, S. Shamail en M.M. Awais, "*Eliminating process of normalization in relational database design*", Proc. INMIC 2003, Pakistan, december 2003.

[KCGS93] W. Kim, I. Choi, S. Gala en M. Scheevel, "*On resolving semantic heterogeneity in multidatabase systems*", IEEE Computer, vol. 24, nr. 12, pp. 12-18, juli 1993.

[LSPR93] E. Lim, J. Srivastava en S. Prabhakar, J. Richardson, "*Entity identification in database integration*", IEEE Computer Society Press, pp. 294-301, 1993.

[LSS96] E. Lim, J. Srivastava en S. Shekhar, "*An evidential reasoning approach to attribute value conflict resolution in database integration*", IEEE, 1996.

[ME96] A.E. Monge en C.P. Elkan, "*The Field Matching Problem: Algorithm and Applications*", Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining, pp. 267-270, augustus 1996.

[PRSL93] S. Prabhakar, J. Richardson, J. Srivastava en E. Lim, "*Instance-Level Integration in Federated Autonomous Databases*", IEEE, 1993.

[SCG91] A.C. Scott, J.E. Clayton, E.L. Gibson, "*A Practical Guide to Knowledge Acquisition*", Addison-Wesley, 1991.

[SKS05] A. Silberschatz, H.F. Korth en S. Sudarshan, *Database System Concepts*, 5th Edition, McGraw-Hill, mei 2005.

[Ste95] M. Stefik, *Introduction to Knowledge Systems*, Morgan Kaufmann, 1995.

[TBDLW87] M. Templeton, D. Brill, S.K. Dao, E. Lund, P. Ward, A.L.P. Chen en R. MacGregor, "*Mermaid – A front end to distributed heterogeneous databases*", Proc. IEEE, vol. 75, pp. 695-708, mei 1987.

[WM89] Y.R. Wang en S. Madnick, "*The interdatabase instance identification problem in integrating autonomous systems*", Proc. 5th Int. Conf. Data Eng., pp. 46-55, februari 1989.

Bijlagen

Naast deze bijlagen horen bij de scriptie de volgende documentbijlagen. De documentbijlagen staan niet in de scriptie, maar zijn los opvraagbaar.

Onderzoeksomgeving – Dit document beschrijft de onderzoeksomgeving. Het document bestaat uit twee onderdelen. Ten eerste wordt de structuur van de onderzoeksomgeving (organisatie, medewerkers) besproken. Ten tweede wordt besproken hoe de expertkennis, die in dit onderzoek gebruikt wordt, verzameld is.

EROS – Dit document bevat alle informatie rond het prototype EROS. Naast de onderdelen uit de scriptie (architectuur, ontwerp en implementatie) bevat het document ook het testplan en documentatie voor gebruik van het prototype.

Bijlage A Informatiebronnen

De inhoud van deze bijlage is grotendeels ontleend aan Criminaliteit en Rechtshandhaving [EH04], een gezamenlijke publicatie van het WODC en het CBS (Centraal Bureau voor de Statistiek). Deze publicatie vormt een jaarlijks geactualiseerd naslagwerk over criminaliteit en rechtshandhaving voor onder meer politiek, beleidsmakers, pers en wetenschap.

A.1 HKS

Het Herkenningsdienstsysteem (HKS) is een landelijk dekkend systeem dat sinds 1986 door de politie gebruikt wordt om gegevens te registeren. Omdat de verschillende politieregio's het systeem beheren, is feitelijk sprake van Herkenningsdienstsystemen: voor elke politieregio en de Koninklijke Marechaussee (KMar) één. De gegevens uit deze deelsystemen worden gecentraliseerd in één systeem, het landelijk HKS. In dit onderzoek wordt het landelijk HKS bedoeld als de term HKS gebruikt wordt.

Het HKS bevat gegevens over 'het begin' van de strafrechtsketen. Het gaat hier om de registratie door de politie van crimineel gedrag (misdrijven). Sinds 1996 is het systeem betrouwbaarder geworden, onder andere doordat alle processen-verbaal, inclusief die van minderjarigen sindsdien moeten worden ingevoerd. Het WODC ontvangt jaarlijks een kopie van het HKS. In dit onderzoek wordt gewerkt met de versie van 2003.

In het HKS wordt in principe elk proces-verbaal opgenomen. Dit betekent natuurlijk niet dat het beeld van criminaliteit compleet is. Daders die niet gepakt zijn, delicten waarvan geen proces-verbaal is opgemaakt, of delicten die niet bekend zijn bij de politie worden niet opgenomen. Voor het onderzoek is dit geen onoverkomelijkheid, aangezien deze delicten ook niet zullen terugkomen in de andere informatiebronnen die later in de strafrechtsketen liggen.

Belangrijker is dat in het HKS alleen natuurlijke personen zijn opgenomen: rechtspersonen zijn uitgesloten van opname. Gegevens van de bijzondere opsporingsdiensten (bijvoorbeeld FIOD, ECD en de douane) zijn uitgezonderd van registratie. Deze gegevens kunnen echter wel tot een zaak leiden en komen dus mogelijk voor in andere informatiebronnen. Verder zijn de politieregio's niet altijd eenduidig in de registratiewijze van een delict. De registratie is bijvoorbeeld mogelijk niet conform indien een verdachte naar een andere (politie)regio verhuist, of een delict pleegt in een (politie)regio waar hij niet woonachtig is. Dit kan mogelijk een rol spelen bij de betrouwbaarheid van attributen.

A.2 OMDATA

OMDATA is een informatiesysteem van het Parket-Generaal (PG) van het Openbaar Ministerie. Het maakt gebruik van een ander, al langer bestaand informatiesysteem, Rapsody genaamd. Rapsody is een gemeenschappelijk informatiesysteem van het Openbaar Ministerie en de zittende magistratuur. Het OMDATA bevat gegevens uit een specifieke module uit Rapsody: de strafrechtmodule (Rapsody strafrechtsketen). Rapsody is op zijn buurt weer gebaseerd op COMPAS, het registratiesysteem dat wordt gebruikt door de arrondissementsparketten en de griffiers van rechtbanken. Rapsody onttrekt bepaalde gegevens aan COMPAS die voor beleidsinformatie van belang zijn, en slaat die op in een gemakkelijk bevraagbare vorm. Het is een decentraal systeem met afzonderlijke databases in elk van de negentien arrondissementen, die ieder alleen gegevens over de strafzaken in het eigen

arrondissement bevatten. Landelijke beleidsinformatie is daardoor niet direct via Rapsody beschikbaar. Die landelijke informatiefunctie wordt sinds midden jaren negentig vervuld door OMDATA.

OMDATA biedt informatie over de instroom van zaken bij het OM, en de afhandeling van die zaken door het OM en door de rechter. OMDATA geeft een volledig beeld van alle afdoeningen[7] (*afgeronde zaken*) sinds 1994. Voor kantonzaken bevat OMDATA tevens het eventuele hoger beroep. Doordat het onderliggende systeem Rapsody in de loop der jaren steeds verder is ontwikkeld, is de informatie in OMDATA over recente zaken gedetailleerder dan die over zaken die in de eerste jaren zijn geregistreerd. Het OMDATA bij de PG wordt elke twee maanden aangevuld met wijzigingen uit Rapsody. Het WODC ontvangt elke vier maanden een kopie van de OMDATA-PG.

OMDATA bevat informatie over de delicten die (primair) ten laste zijn gelegd. Hiermee ligt OMDATA dus dichtbij de gegevens van het HKS, waarin de geconstateerde delicten in een proces-verbaal genoemd worden. De laatste jaren bevat OMDATA ook informatie over de delicten waarvoor is veroordeeld. Deze informatie is ook in de OBJD terug te vinden.

Naast OMDATA bevat ook de OBJD gegevens over strafzaken die afkomstig zijn uit COMPAS. Er zijn echter enkele belangrijke verschillen tussen deze informatiebronnen. De OBJD is gebaseerd op de justitiële documentatie en geeft informatie over onherroepelijke afdoeningen. Bij OMDATA blijft de informatie beperkt tot de afdoening in eerste aanleg. Bovendien is de informatie in de OBJD op persoonsniveau beschikbaar (zij het geanonimiseerd), terwijl OMDATA alleen informatie op zaakniveau biedt.

A.3 OBJD

In de jaren 1992-1996 zijn de Algemene Documentatieregisters, waarin de zogenaamde ‘strafbladen’ werden bijgehouden, geautomatiseerd en samengevoegd in een landelijk systeem, het Justitieel Documentatie Systeem (JDS). De Onderzoeks- en Beleidsdatabase Justitiële Documentatie (OBJD) bevat een selectie uit en soms een bewerking van de gegevens uit het JDS. De gegevens in de OBJD zijn geanonimiseerde, persoonsgebonden gegevens over aard en onherroepelijke afdoening van alle misdrijven en een aantal overtredingen, die na 1996 door de rechtelijke macht in behandeling zijn genomen. Het bevat ook een aantal persoonsgegevens, kennisgevingen van gebeurtenissen na het vonnis (bijvoorbeeld tenuitvoerlegging of het verstrijken van de proeftijd) en gegevens over buitenlandse strafzaken. De plegers kunnen zowel natuurlijke personen als rechtspersonen zijn. Van alle personen die na 1996 in aanraking met justitie komen, wordt de justitiële voorgeschiedenis ingevoerd. Dit kan teruggaan tot het begin van de vorige eeuw.

De OBJD is pas sinds 1996 volledig, in de jaren daarvoor bevat zij alleen gegevens over personen en hun strafzaken als er ook na 1996 een justitieel contact is geweest. De OBJD heeft de gegevens over de onherroepelijke afdoeningen. Dit betekent dat vonnissen in eerste aanleg alleen opgenomen worden als dit vonnis onherroepelijk is geworden. Als er in hoger beroep is ingesteld, blijft de zaak in de OBJD een openstaande zaak, tot het vonnis in hoger beroep onherroepelijk is geworden. De gegevens over deze laatste uitspraak worden dan geregistreerd, niet die van het vonnis in eerste aanleg.

[7] Een afdoening is een afgeronde zaak, door bijvoorbeeld strafoplegging of (geld)transactie.

Bijlage B Literatuuronderzoek

Centraal in dit onderzoek staat de reconciliatie van objectgelijke entiteiten. Om dit te bereiken moeten de entiteiten met elkaar vergeleken worden. Dit impliceert:
- De entiteiten moeten in elke bron uniek geïdentificeerd kunnen worden.
- De verbanden tussen entiteittypes in de verschillende bronnen moeten gebruikt worden om entiteiten aan elkaar te koppelen.

Om entiteiten uniek te identificeren, moet een entiteittype een uniek attribuut hebben, of moet een identificerende set van attributen gebruikt kunnen worden. Wanneer entiteiten fysiek niet als zodanig zijn opgeslagen, dan kan het in het laatste geval voorkomen dat verschillende representaties of inconsistenties ervoor zorgen dat een entiteit dubbel geïdentificeerd wordt. In deze opdracht wordt er vanuit gegaan dat dit vooraf opgelost is of voor lief wordt genomen.

B.1 Overeenkomsten tussen informatiebronnen

De overeenkomsten die bestaan tussen entiteittypen in de verschillende informatiebronnen vormen de enige mogelijkheid om entiteiten aan elkaar te koppelen. Het selecteren van de attributen die gemeenschappelijke eigenschappen beschrijven heet *attribuutselectie*.

In dit stadium is het goed om op te merken dat bij dit onderwerp (koppelen van objectgelijke entiteiten) vaak wordt uitgaan van één entiteittype. In dit geval betreft het volledige informatiebronnen met meerdere entiteittypes, waarbij het niet noodzakelijk zo hoeft te zijn dat entiteittypes in een eigen tabel staan. De entiteittypes en zelfs hun attributen moeten dus nog bij elkaar gezocht worden. Naast attribuutselectie is er in deze opdracht dus ook *schema-integratie* nodig.

Als elementen in twee verschillende informatiebronnen gekoppeld worden, die verwijzen naar hetzelfde element in de reële wereld, dan heet dit *reconciliatie*. De elementen worden door de koppeling weer bij elkaar gebracht. Reconciliatie kan toegepast worden op entiteiten en attributen. Gemeenschappelijke entiteittypen en informatiemodellen worden *geconcilieerd*, waarbij de onderliggende entiteiten en eventueel attributen gereconcilieerd worden.

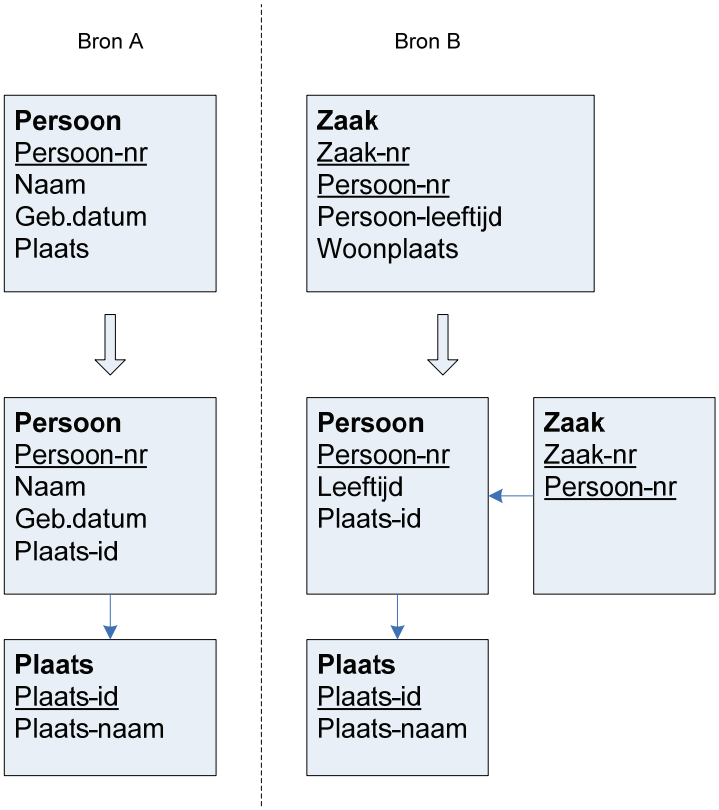
B.2 Schema-integratie

Schema-integratie is het integreren van schema's van verschillende informatiebronnen. Uiteindelijk kunnen hiermee de gegevens in de informatiebronnen zelf worden samengevoegd (data-integratie), of kan er een schil op de bronnen gelegd worden, waardoor zij op een zelfde manier te benaderen zijn.

De meeste studies op het gebied van schema-integratie gaan uit van gemeenschappelijke sleutels. Hierdoor kan de koppeling op entiteitniveau al plaatsvinden. Problemen die onderzocht worden zijn dan inconsistenties [AKWS95], detecteren van semantisch gelijke attributen [KCGS93, DeM89] en conflicten bij attribuutintegratie [LSS96]. In dit onderzoek ligt de nadruk op het identificeren van objectgelijke entiteiten [Dey02, WM89].

Het doel van schema-integratie is hier het vinden van gezamenlijke entiteittypes en hun attributen. Hierdoor moet het mogelijk worden een gemeenschappelijk datamodel te genereren uit de bestaande schema's van de informatiebronnen [HSA03]. Dit heeft een aantal voordelen:

- Het koppelproces kan zijn werk efficiënter doen.
- Het wordt eenvoudiger een gemeenschappelijke informatiebron op te zetten waarbij
    - attributen worden samengevoegd;
    - de structuur voor bevragers logisch is opgebouwd.
- De betekenis van attributen wordt gedocumenteerd (gebruikt bij attribuutselectie).



Figuur Bijlage B.1 – Omvorming van schema's naar een gemeenschappelijk datamodel.

Vervolgens kunnen de objecten Plaats en Persoon worden geconcilieerd. Er zijn verschillende strategieën op het gebied van schema-integratie. Deze overlappen vaak met de strategieën van attribuutselectie [PRSL93]. Deze worden daarom samen besproken.

B.3 Attribuutselectie

Wanneer de entiteittypes zijn vastgesteld, kunnen de attributen van entiteittypes gebruikt worden in het koppelproces. Het helpt hierbij om de attributen te *centraliseren* rond een entiteittype, bijvoorbeeld Persoon (in bron B, zie boven voor het model):

Figuur Bijlage B.2

Merk op dat de zaken van een persoon als één attribuut (set van zaken) onder Persoon kan worden gezien. Hierdoor kunnen eigenschappen uit de volledige informatiebron ingezet worden bij het koppelen van een objectgelijke persoonsentiteiten. Tussen attributen kunnen verschillende overeenkomsten of verbanden worden vastgesteld. Op basis hiervan zijn in het verleden ook verschillende methodes ontworpen.

B.3.1 Gelijke definitie

Dit lijkt de meest logische keuze: attributen met elkaar in verband brengen als zij hetzelfde gedefinieerd zijn. Dit is op meerdere manieren uit te leggen: attributen hebben dezelfde naam, of attributen hebben dezelfde betekenis (domein, semantiek) [CER87], [BOT86], [TBDLW87]. In beide gevallen gelden dezelfde positieve en negatieve eigenschappen.

Positief	Omschrijving
Snelheid	Het is op deze manier snel duidelijk welke attributen bij elkaar horen. Voor een expert zeker, voor een algoritme moeilijk (zie automatisering).
Automatisering	Zonder extra gebruikersinformatie kunnen attributen met elkaar in verband worden gebracht door een algoritme. Uit eerdere studies [KCGS93] is wel bekend dat dit problemen oplevert, bijvoorbeeld door andere naamgeving. Ontology-databases [BDS03] geven hiervoor een (beperkte) oplossing.

Tabel Bijlage B.1 – Gelijke definitie – Voordelen

Negatief	Omschrijving
Foute selectie	Met naamgeving kan niet alles worden weergeven en daarom is het soms triviaal: het attribuut *hoogte* van een gebouw is niet hetzelfde als het attribuut *hoogte* van een persoon. Zo zijn er nog meer voorbeelden te noemen.
Incompleet	Door alleen te focussen op attributen met dezelfde definitie, worden andere verbanden opzij geschoven. Zo bestaat er een sterk verband tussen leeftijd en geboortedatum. Dit soort verbanden spelen in deze opdracht een grote rol.

Tabel Bijlage B.2 – Gelijke definitie – Nadelen

B.3.2 Gelijke inhoud

Methodes die entiteiten proberen te koppelen werken meestal op één tabel, met enkele attributen. Door het kleine aantal attributen kan het haalbaar zijn om uit te zoeken welke attributen inhoudelijk het meeste met elkaar te maken hebben [DSD02]. Zo kan een algoritme bijvoorbeeld ontdekken dat *Straat* en *Huisnummer* in bron A vergelijkbaar zijn met *Adres* in bron B.

Positief	Omschrijving
Objectief	Ongeacht de brondefinities worden attributen vergeleken aan de hand van de inhoud. Dit levert een objectief oordeel op over de mate van overeenkomst.
Geen problemen met definities	De brondefinities hoeven niet nauwkeurig bepaald te worden, omdat hier in geen tot mindere mate naar gekeken wordt.

Tabel Bijlage B.3 – Gelijke inhoud – Voordelen

Negatief	Omschrijving
Tijdrovend	Omdat alle gegevens onderzocht moeten worden, is het bepalen van verbanden tussen attributen een tijdrovend proces.
Risico op verkeerde verbanden	Deze methode vergelijkt ook attributen met verschillende definities. Hierdoor bestaat een nog grotere kans dan eerder, dat attributen onterecht met elkaar in verband worden gebracht. Zo kan nu ook *Woonplaats* met *Plaats delict* in verband worden gebracht.

Tabel Bijlage B.4 – Gelijke inhoud – Nadelen

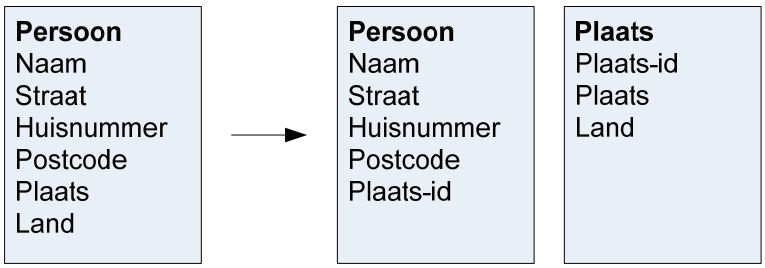
B.3.3 Overige overeenkomsten

Bovenstaande overeenkomsten zijn gebaseerd op naamgeving, domein of inhoud. Er zijn echter ook overeenkomsten die niet direct uit de bronnen afgeleid kunnen worden:

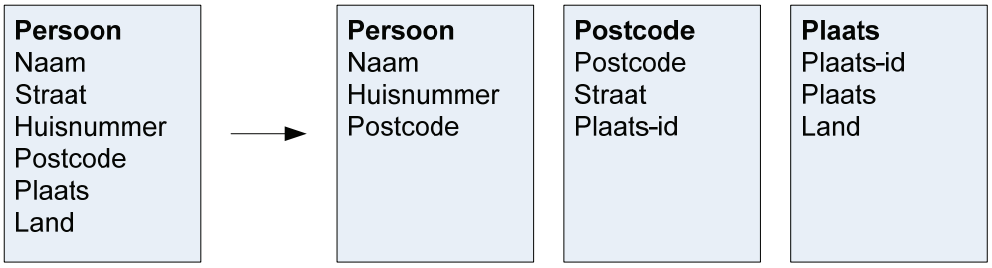
- Functionele afhankelijkheden.

  Bijvoorbeeld: als een persoon is veroordeeld tot gevangenisstraf, dan is de lengte van de gevangenisstraf ook bekend.

  Functionele afhankelijkheden kunnen gebruikt worden bij het maken van een genormaliseerd relationeel model. Automatisch normaliseren kan tot een aantal problemen leiden:
  - niet alle functionele afhankelijkheden hoeven genormaliseerd worden:



Figuur Bijlage B.3 – Voorbeeld van gematigde normalisering



Figuur Bijlage B.4 – Voorbeeld van extreme normalisering

  - functionele afhankelijkheden worden afgeleid door inductie, waardoor de geldigheid niet absoluut zeker is:

Dataset 1				Dataset 2		
ID	*Waarde*	*Tekst*		*ID*	*Waarde*	*Tekst*
1	1	a		4	1	a
2	2	b		5	2	b
3	1	a		6	3	a

Tabel Bijlage B.5 – Functionele afhankelijkheden

`Waarde` → `Tekst` is in de linker dataset nog een afhankelijkheid, maar in de rechter dataset blijkt dit niet te kloppen.

- Condities.

  Bijvoorbeeld: als het *Peiljaar* van een delict na 2000 is, dan is de *Nationaliteit ouders* ingevuld.

- Relaties tussen bronnen.

  Bijvoorbeeld: Als een delict in bron A een misdrijf is, dan komt het vrijwel zeker ook in bron B voor (omgekeerd hoeft niet te gelden).

- Onderzoeksverbanden.

Met onderzoeksverbanden (correlaties) worden verbanden bedoeld die niet gebruikt kunnen worden in de koppeling, omdat ze niet volgen uit vaststaande feiten. Bijvoorbeeld: personen die een vermogensdelict begaan kunnen gemiddeld veel ouder zijn dan personen die andere delicten plegen. Dit is een patroon dat geverifieerd kan worden nadat de koppeling is gemaakt, maar dat in eerste instantie niet gebruikt mag worden in de koppeling. Na verificatie kan het verband eventueel gebruikt worden.

Bijlage C Theoretische achtergrond

C.1 Gewicht in kennisregels

In het Expert Kennis Systeem hebben experts de mogelijkheid een gewicht op te geven bij een kennisregel. Dit gewicht is het initieel gewicht. Als de expert hier geen gebruik van maakt, dan wordt het initieel gewicht berekend. Dit wordt gedaan aan de hand van de discriminerende attributen $attr_1$ en $attr_2$.

Onder discriminatie wordt verstaan: de mate waarin de attributen in staat zijn entiteiten te identificeren. Hierbij zijn twee elementen te onderscheiden:

1. Hoe discriminerend zijn de attributen in hun eigen informatiebronnen? Dit wordt het *initieel gewicht* (w_0) genoemd.
2. Hoe discriminerend zijn de attributen in het reconciliatieproces? Dit wordt het *gewicht bijstellen* genoemd.

Definitie. Gegeven een kennisregel met een door de expert opgegeven gewicht u en de mate van discriminatie $id(attr_1)$ en $id(attr_2)$ van respectievelijk het attribuut uit bron I en bron II. Het initiële gewicht w voor de kennisregel wordt gedefinieerd als:

$$w_0 = \begin{cases} u & \exists u \\ \dfrac{id(attr_1) + id(attr_2)}{2} & \neg\exists u \end{cases}$$

De mate waarin een attribuut discriminerend is, wordt berekend uit de verhouding tussen het aantal unieke waarden en het aantal totale waarden. De waarde NULL wordt in deze berekening niet meegenomen. Zo worden attributen beoordeeld op hun mate van discriminatie voor de waarden waarvoor ze ingevuld zijn.

Voorbeeld. GBA-nummer in OBJD is uniek identificerend wanneer deze is ingevuld. In andere gevallen is het GBA-nummer leeg en levert de vergelijking *n.a.* op, waardoor het zware gewicht niet meetelt in de uiteindelijke berekening.

Definitie. Gegeven een attribuut a en de volgende gegevens over het domein van a:

$d(a)$ aantal unieke waarden
$e(a)$ aantal lege waarden
$s(a)$ totaal aantal waarden

De mate van discriminatie $id(a)$ wordt gedefinieerd als:

$$id(a) = \begin{cases} 0 & \text{als } s(a) = 0 \vee s(a) = e(a) \\ \dfrac{d(a)}{s(a)} & \text{als } e(a) = 0 \wedge s(a) \neq e(a) \\ \dfrac{d(a) - 1}{s(a) - e(a)} & \text{als } e(a) \neq 0 \wedge s(a) \neq e(a) \end{cases}$$

Als het bijstellen van het gewicht geactiveerd is, dan wordt tijdens het reconciliatieproces de uitkomst van de kennisregel voortdurend vergeleken met andere uitkomsten. Hierbij wordt het

gewicht bijgesteld door het aandeel van de kennisregel in de totale gelijkheid te waarderen. Bijvoorbeeld:

Stel, de vergelijking van een entiteitcombinatie *i* levert de gelijkheid sim_i op. Hieraan heeft de gelijkheid van kennisregel *j* (r_j) bijgedragen. Dit aandeel moet hoog gewaardeerd worden als de totale gelijkheid hoog is, en laag als de totale gelijkheid ook laag is. Een hoge gelijkheid van de kennisregel bij een lage totale gelijkheid moet echter weer laag gewaardeerd worden. De vergelijking $sim_i \cdot r_j$ voldoet aan deze eisen.

Het gewicht kan bijgesteld worden door de gemiddelde waardering van het aandeel van de kennisregel mee te nemen (n = volgnummer van de huidige vergelijking, j = nummer van de kennisregel).

Definitie. Het bijgestelde gewicht wordt gedefinieerd als:

$$w_n = \begin{cases} w_0 & \text{als } n = 1 \\ \dfrac{w_0}{n-1} \sum_{i=1}^{n-1} sim_i \cdot r_j & \text{als } n > 1 \end{cases}$$

Bij het toepassen van het bijgestelde gewicht zijn enkele kritische kanttekeningen te plaatsen:

- Het bijgestelde gewicht is niet direct betrouwbaar. Pas nadat er zoveel vergelijkingen zijn gedaan dat een representatieve set van entiteiten vergeleken is, is het gewicht echt betrouwbaar. Er kan besloten worden om het bijgestelde gewicht pas in te zetten na een aantal vergelijkingen, maar de betrouwbaarheidsgrens kan niet met zekerheid worden bepaald. Een mogelijkheid is het bijgestelde gewicht pas na het proces te gebruiken, door het te vergelijken met het initieel gewicht en daar conclusies uit te trekken.
- Het bijstellen van gewicht kan alleen als er voldoende andere kennisregels zijn om een kennisregel mee te vergelijken. Bij te weinig kennisregels wordt de totale gelijkheid te veel bepaald door de gelijkheid die de kennisregel zelf oplevert, of door een enkele kennisregel met een relatief groot gewicht.

In dit onderzoek is het bijstellen van het gewicht uiteindelijk niet gebruikt vanwege de tweede kanttekening: een gebrek aan kennisregels.

C.2 Clustering

In de theorie wordt clustering gebruikt om het aantal vergelijkingen te verminderen. Clustering is bedoeld om de efficiëntie te verhogen; de resultaten veranderen er niet door. Clustering is dan ook geen noodzaak, de balans tussen tijdswinst in het proces en complexiteit (en daarmee implementatietijd) is belangrijk. Deze paragraaf laat zien dat:

- teveel clustering averechts werkt;
- bij clustering door middel van één sleutel alle vormen van clustering mogelijk zijn.

Omdat teveel clustering averechts werkt, is het niet wenselijk om te zoeken naar maximale clustering. Het zoeken naar de optimale clustering vergt veel rekenkracht en een veel ingewikkelder ontwerp van het prototype. Aangezien clustering in de casus toch vaak slechts met één ouder mogelijk is, én omdat clustering met meer ouders indien nodig toch gebeurt door middel van één clustersleutel (en dus uiteindelijk één waarde), is gekozen voor clustering met één waarde. Bijkomend voordeel is dat het ontwerp en de implementatie van het prototype

hierdoor eenvoudiger is, maar eventueel ook eenvoudig uit te breiden naar clustering met meer waarden.

Teveel clustering werkt averechts

In de vergelijking voor objectgelijkheid komt de oudergelijkheid voor entiteitgelijkheid. Hierdoor wordt gelijkheid al niet onnodig berekend; het aantal uiteindelijke berekeningen is gelijk. De winst van clustering zit dus puur in het verminderen van het aantal entiteiten dat met elkaar vergeleken wordt. Om dit aantal te verminderen, worden de entiteiten in clusters ingedeeld en vergeleken.

Hierbij spelen twee factoren een rol: de tijd die het kost om een recordset te openen (de kosten van een *query*, c_q) en de tijd die het kost om een record op te halen (de kosten van een *fetch*, c_f). Een cluster i heeft in elke bron een aantal entiteiten: $e_{1,i}$ en $e_{2,i}$ voor respectievelijk bron 1 en bron 2. Bij het verwerken van een cluster wordt de query geopend met entiteiten uit bron 1, waarbij alle entiteiten worden afgelopen. Voor elk record uit bron 1 wordt de query geopend met entiteiten uit bron 2 en deze entiteiten worden vergeleken met de huidige entiteit in de eerste query. Er geldt:

$$c = p \cdot c_q + \sum_{i=1}^{p}(e_{1,i} \cdot (c_f + c_q + e_{2,i} \cdot c_f)) = c_q \cdot \sum_{i=1}^{p}(1 + e_{1,i}) + c_f \cdot \sum_{i=1}^{p}(e_{1,i} \cdot (1 + e_{2,i}))$$

Het openen van een query vergt meer rekenkracht dan het ophalen van een record ($c_q > c_f$). Er is daarom een break-even punt in bovenstaande kostenformule. Omdat het aantal entiteiten per cluster verschillend is, is het berekenen van dit punt een zware operatie. Om een indruk te geven, kan de situatie worden geschat. Stel de bronnen bevatten evenveel entiteiten ($e_1=e_2$, aangeduid met e) en de clustering verdeelt de bronnen in gelijke partities ($\forall i \in \{1,2\} \forall j \in \{1..p\} : e_{i,j}=e/p$). De formule voor het optimale aantal clusters p in het break-even punt kan nu in een aantal stappen berekend worden:

$$c_q \cdot (p + \frac{e}{p}) - c_f \cdot (\frac{e + e^2}{p}) = 0$$

$$c_q \cdot p + \frac{c_q \cdot e}{p} - \frac{c_f \cdot e - c_f \cdot e^2}{p} = 0$$

$$c_q \cdot p^2 + c_q \cdot e - c_f \cdot e - c_f \cdot e^2 = 0$$

$$c_q \cdot p^2 = c_f \cdot e^2 + (c_f - c_q) \cdot e$$

$$p = \sqrt{(c_f \cdot e^2 + (c_f - c_q) \cdot e) / c_q}$$

Het aantal clusters p ligt tussen 1 en $e_1 \cdot e_2$. Uit de bovenstaande formules is af te leiden dat teveel clustering een averechts effect heeft, aangezien bij een grote p de kosten c_q meer invloed gaan hebben dan de kosten c_f.

Verder is uit de formules af te leiden dat het optimale aantal clusters toeneemt als het aantal entiteiten toeneemt. Zonder de eerder gemaakte aannames ($e_1=e_2$, $\forall i \in \{1,2\} \forall j \in \{1..p\} : e_{i,j}=e/p$) kan de exacte optimale waarde van p alleen berekend worden als alle aantallen per cluster voor elke waarde van p worden berekend. Het is duidelijk dat dit een extreem zware berekening is, waarvoor meer onderzoek nodig is om dit efficiënt te kunnen doen.

Clustering door middel van een clustersleutel

In de theorie van dit onderzoek is sprake van clustering met behulp van één waarde: een ouderknoop met een één-op-meer relatie met de kindknoop. Bij clustering met één ouderknoop heeft elke entiteit één ouder die de clustering bepaalt. De gelijkheidswaarde van deze ouder wordt gedefinieerd als de *clustersleutel*. Bij meerdere ouderknopen is er sprake van meerdere gelijkheidswaarden. De clustersleutel kan nu geschreven worden als een $1\times n$-matrix A, waarbij n het aantal ouderknopen is. Entiteiten met dezelfde clustersleutel hoeven alleen berekend te worden als geldt: $A_{1,1} \circ \ldots \circ A_{1,n} \neq 0$.

Bij meer-op-meer relaties is de clustersleutel geen enkele waarde of een set van waarden met vaste lengte. In het geval van één meer-op-meer relatie is de clustersleutel een set van waarden met variabele lengte: een $m\times 1$-matrix. In het geval van meerdere meer-op-meer relaties is de clustersleutel zelfs een $m\times n$-matrix met variabele m en vaste n. Entiteiten met dezelfde clustersleutel hoeven alleen berekend te worden als geldt: $A_{1,1} \circ \ldots \circ A_{m,n} \neq 0$.

Bij clustering met behulp van één ouderknoop bestaat de clustersleutel uit één gelijkheidswaarde. Elke andere clustering bestaat de clustering uit een matrix van gelijkheidswaarden. De clustersleutel is uniek, de entiteitcombinaties waaruit de gelijkheids-waarden berekend worden niet. Om dubbele berekeningen te voorkomen, moeten de clustersleutels vooraf berekend worden. Door deze uitkomsten in een tabel te plaatsen, kan deze tabel als 'clusterouder' van de knoop optreden. De clustersleutel vormt de primaire sleutel van de ouder en de gelijkheidswaarde is gelijk aan $A_{1,1} \circ \ldots \circ A_{m,n}$.

Doordat de clustersleutels berekend moeten worden voordat de clustering wordt toegepast, staan de bijbehorende knopen in een ander informatiemodel. Het genereren van de clustersleutels met de bijbehorende gelijkheidswaarde is de schakel tussen het 'ouder'-informatiemodel en het 'kind'-informatiemodel waarin de clustering wordt toegepast. De implementatie van deze schakel wordt overgelaten aan toekomstig werk.

C.3 Modelconciliatie

De theorie in dit onderzoek streeft ernaar om de objectgelijkheid van de centrumknoop zo goed mogelijk te beschrijven. Doordat de gelijkheid wordt gedistribueerd door de andere knopen, heeft elke knoop zijn eigen objectgelijkheid. Deze objectgelijkheid mist alleen de gelijkheid uit de richting van de centrumknoop. Deze paragraaf beschrijft hoe te werk gegaan moet worden om alle knopen in een informatiemodel optimaal te conciliëren.

Ouderknopen

Door hun positie in de hiërarchische structuur wordt de gelijkheid in ouderknopen eerder berekend dan in de centrumknoop. Dit betekent dat een informatiemodel met twee hiërarchische modellen hetzelfde resultaat oplevert.

Voorbeeld. Gegeven de volgende gelijkheidsstelsels van informatiemodellen, in vereenvoudigde notatie:

$$S_1 = \{\underline{c} = (a)c$$

$$S_2 = \begin{cases} \underline{a} = a \\ \underline{c} = \underline{a}c \end{cases}$$

Er geldt dat S_1=S_2, de objectgelijkheid is in beide stelsels hetzelfde. Het tweede stelsel biedt echter wel een nieuwe mogelijkheid. Als *A* zonder de centrumknoop *C* niet voldoende beschreven wordt, dan kan deze knoop in het tweede stelsel worden toegevoegd:

$$S_3 = \begin{cases} \underline{a} = a(c) \\ \underline{c} = \underline{a}c \end{cases}$$

Daarnaast biedt een gesplitst stelsel nog een voordeel. Na de conciliatie van het eerste informatiemodel kunnen de reconciliaties bevestigd worden. In plaats van gelijkheidswaarden leveren de vergelijkingen dan *0* of *1*. De clustering is hierdoor optimaal. De aanbeveling voor ouderknopen luidt dan ook om ze vooraf in een eigen informatiemodel te conciliëren en de reconciliaties te bevestigingen voordat het volgende informatiemodel geconcilieerd wordt.

Kindknopen

De kindknopen vormen in het informatiemodel een eigen hiërarchisch model onder de centrumknoop. Alleen de gelijkheid uit centrumknoop wordt hierin niet meegenomen. Echter, door de clustering wordt er op kindniveau in feite vanuit gegaan dat de ouders gelijk zijn. Deze gelijkheid telt alleen niet mee in de objectgelijkheid. Als de eigenschappen van een kindknoop sterk genoeg zijn, dan zijn de reconciliaties van de kindknoop na de conciliatie van de centrumknoop ook geldig. Als de eigenschappen niet sterk genoeg zijn, dan moet de kindknoop na afloop in een eigen model opnieuw geconcilieerd worden. De centrumknoop treedt dan op als ouderknoop.

Conclusie

Voor een optimale conciliatie van een volledig hiërarchisch model, kan het hiërarchisch model het beste van boven naar beneden en knoop voor knoop geconcilieerd worden.

C.4 Samenvoegen gelijkheid

De samenvoegfunctie voor gereconcilieerde gelijkheid $m(K)$ berekent, zoals besproken in de theorie, een gelijkheidswaarde uit de gelijkheidsmatrix uit de gelijkheid van gereconcilieerde elementen. Hier worden enkele alternatieven kort besproken.

Penalty voor ongereconcilieerde gelijkheid

Gegeven de definities zoals beschreven in §3.4. De samenvoegfunctie $m_p(K)$ berekent de gelijkheidswaarde uit gereconcilieerde elementen, maar verlaagt de gelijkheid in geval van ongereconcilieerde gelijkheid:

$$m_p(K) = m(K) \cdot \frac{\sum(SM_{i,j} \cdot occ(i,j) \cdot recon(i,j))}{\sum(SM_{i,j} \cdot occ(i,j))} = \frac{m(K)^2}{\sum SM_{i,j} \cdot occ(i,j)}$$

Andere methoden

Eerdere methodes werken op basis van reconciliatie. Dit is een zwaar proces, wat wellicht niet altijd nodig is of zelfs niet het beste werkt. Soms kan het genoeg zijn om te kijken naar de gelijkheidswaarden zelf. Gegeven de volgende gelijkheidsmatrix:

	A	B	C	D
1	1	0,2	-	-
2	-	-	0,6	-
3	-	-	0,3	0,5
4	-	0,8	-	-

Tabel Bijlage C.1

Over deze matrix kunnen enkele statistieken berekend worden, zoals de minimum gelijkheid en maximum gelijkheid van de rijen en kolommen:

	A	B	C	D	row_{min}	row_{max}
1	1	0,2	-	-	0,2	1
2	-	-	0,6	-	0,6	0,6
3	-	-	0,3	0,5	0,3	0,5
4	-	0,8	-	-	0,8	0,8
col_{min}	1	0,2	0,3	0,5		
col_{max}	1	0,8	0,6	0,5		

Tabel Bijlage C.2

Er geldt nu verder:

$$\min(col_{\min}) = \min(row_{\min}) = \min(matrix) = 0{,}2$$
$$\max(col_{\max}) = \max(row_{\max}) = \max(matrix) = 1$$

$$\min(row_{\max}) = 0{,}5$$
$$\max(row_{\min}) = 0{,}8$$

$$\min(col_{\max}) = 0{,}5$$
$$\max(col_{\min}) = 0{,}5$$

Al deze getallen kunnen in specifieke gevallen de gelijkheid beschrijven, zoals in Tabel Bijlage C.3 staat aangegeven.

Getal	**Omschrijving**
min(matrix)	Alle elementen uit bron I moeten goed scoren met alle elementen uit bron II.
max(matrix)	In elk geval één element uit bron I moet goed scoren met in elk geval één element uit bron II.
$min(row_{max})$	Alle elementen uit bron I moeten in elk geval éénmaal goed scoren met een element uit bron II.
$max(row_{min})$	In elk geval één element uit bron I moet goed scoren met alle elementen uit bron II.
$min(col_{max})$	Zie $min(row_{max})$, maar dan vanuit bron II.
$max(col_{min})$	Zie $max(row_{min})$, maar dan vanuit bron II.

Tabel Bijlage C.3

Bijlage D Voorbereiding casus

Dit onderzoek richt zich op de reconciliatie van persoonsentiteiten. Voordat het informatiemodel van Persoon geconcilieerd kan worden, moeten de informatiemodellen van Land en Wetsartikel geconcilieerd worden. Deze conciliaties worden besproken in deze bijlage.

D.1 Reconciliatie landen

Deze paragraaf bespreekt hoe de landen (in dit geval geboortelanden) gereconcilieerd zijn. De geboortelanden verwijzen over het algemeen naar de landindeling op het moment van invoeren, en kan daarom per persoon verschillen. Bij het maken van een landsindeling spelen een aantal zaken een rol:

- Wat is de definitie van een land?

  Zowel de Sovjetunie als Wit-Rusland zijn landen (geweest). Maar ook de Nederlandse Antillen worden als apart land behandeld door de informatiebronnen.

- Hoe verhoudt het land zich tot andere landen?

  Is het land ooit onderdeel geweest van een ander land? Of bevat het nu landen die vroeger zelfstandig waren?

Synoniemen voor hetzelfde land hebben gelijkheid 1. Voor een land dat ontstaan is nadat het moederland uiteen is gevallen, kan de gelijkheid gelijk gesteld worden aan de kans dat een persoon met geboorteland het moederland, uit het ontstane land afkomstig is.

Omdat in de casus de nadruk wordt gelegd op vergelijkingen van processen-verbaal, worden aan persoonskenmerken hoge eisen gesteld: een reconciliatiemogelijkheid wordt afgekeurd als het geboorteland niet hetzelfde is. Om te voorkomen dat reconciliatiemogelijkheden ten onrechte worden afgekeurd, zijn een aantal landen samengenomen (bijvoorbeeld Zaïre, Congo, Kongo-Kinshasa en Kongo-Brazzaville). Uiteindelijk is 1,3% niet gekoppeld omdat de geboortelanden niet overeenkwamen. Hiervan had 56% potentieel gereconcilieerd kunnen worden als er meer landen zouden worden samengevoegd. Dit zou echter wel nadelige invloed hebben op de waarschijnlijkheid van reconciliatiemogelijkheden.

Er is daarom voor de volgende indeling gekozen:

- Indien het kleine landen betreft, of de aantallen in de informatiebronnen niet significant zijn, dan worden deze landen samengevoegd tot één land (meestal het grootste).
- Koloniale gebieden: in sommige gevallen worden huidige koloniale gebieden samengevoegd tot één land (onder het moederland), omdat deze geografisch ver van het moederland verwijderd zijn (verhouding 9:1). Oude koloniale gebieden besloegen veelal meerdere landen. Deze gebieden verwijzen nu slechts naar het moederland.
- Indien het grote landen betreft (uiteenvallen van de Sovjetunie), of samenvoegen een grote impact zou hebben op de reconciliaties (Nederlandse Antillen, Suriname), dan wordt de vergelijking in tweeën gesplitst (verhouding 4:1).

De landen worden dus ingedeeld in drie onderdelen:
- moederland;
- land (verhouding 4:1 ten opzichte van moederland);
- gebied (verhouding 9:1 ten opzichte van moederland).

Land en gebied zijn niet met elkaar te vergelijken, omdat de een geen deelverzameling van de ander is. Mocht dit voorkomen, dan wordt op moederland vergeleken en de minimale gelijkheid opgeleverd (80%).

Bijvoorbeeld:
- Tsjechië en Slowakije worden samengevoegd onder Tsjecho-Slowakije.
- De Cocoseilanden en Christmaseiland vallen onder Australië (90%) en de Australische eilanden (10%).
- Wit-Rusland valt onder Sovjetunie (80%) en Wit-Rusland (20%).

Zodoende resulteert een vergelijking tussen Tsjechië en Slowakije in 1 en een vergelijking tussen Wit-Rusland en de Sovjetunie in 0.8.

Andere manieren om de overeenkomsten tussen landen te beschrijven zijn zeker mogelijk. Met de gekozen manier wordt een bevredigend resultaat behaald en dat is – met het oog op de beschikbare tijd – voldoende voor dit onderzoek.

Bij de reconciliatie van landen is de volgende gelijkheidsfunctie gebruikt[8]:

$$sim_{land} = \begin{cases} 1 & \text{als } land_1 = land_2 \\ 0.9 & \text{als } land_1 = land_2 \wedge kleingrondgebied_1 \neq kleingrondgebied_2 \\ 0.8 & \text{als } land_1 = land_2 \wedge grootgrondgebied_1 \neq grootgrondgebied_2 \\ 0.5 & \text{als } land_x = \text{V.A.R.} \wedge land_y \in (\text{Egypte}, \text{Syrië}) \\ 0.5 & \text{als } land_x = \text{Duits Oost} - \text{Afrika} \wedge land_y \in (\text{Burundi}, \text{Rwanda}, \text{Tanganyika}) \\ 0.3 & \text{als } land_x = \text{Oost Afrika} \wedge land_y \in (\text{Kenia}, \text{Oeganda}, \text{Tanzania}) \\ 0.3 & \text{als } land_x = \text{Oostenrijk - Hongarije} \wedge land_y \in (\text{Oostenrijk}, \text{Hongarije}) \\ 0 & \text{anders} \end{cases}$$

In deze formule geldt: $x \in \{1,2\} \wedge y \in \{1,2\} \wedge x \neq y$

D.2 Reconciliatie wetsartikelen

De vergelijking van wetsartikelen vormt een belangrijk onderdeel van het reconciliatieproces. Door de wetsartikelen te reconciliëren voor gebruik in de persoonsreconciliatie is deze vergelijking aanzienlijk versterkt. De wetsartikelen werden in HKS en OMDATA verschillend gerepresenteerd. Het reconciliëren van wetsartikelen bestond hoofdzakelijk uit het gelijktrekken van de verschillende representaties. Dit proces wordt als eerste beschreven. Daarna wordt dieper ingegaan op de gelijkheidswaarde die verschillende wetsartikelen ten opzichte van elkaar kunnen hebben, doordat ze mogelijk hetzelfde delict kunnen beschrijven.

[8] V.A.R. staat voor Verenigde Arabische Republiek

D.2.1 Representatie

Het eerste dat opvalt als de wetsartikelen worden vergeleken tussen HKS en OMDATA, is het verschil in scheidingsteken. In HKS wordt over het algemeen een punt als scheidingsteken gebruikt, in OMDATA een slash (SR 280.1 versus SR 280/1). Dit is echter niet consequent doorgevoerd. Zo worden onderdelen van artikelen op verschillende manieren gerepresenteerd:

- SR 250A/1/1°
- SR 250TER/1/1
- ARBTBV 2.4:2/1

Verder is bekend dat de wetboeken op verschillende manieren worden afgekort. Zo is de opiumwet OPIUMW in HKS en OW in OMDATA. Om de wetsartikelen te kunnen reconciliëren moeten eerst dit probleem worden opgelost.

Wetboeken

De meeste afkortingen voor wetboeken kwamen overeen. Door expertkennis en het vergelijken van artikelnummers zijn de volgende wetboeken gevonden die meerdere afkortingen hadden:

Wetboek	Afkortingen
Afvalstoffenwet	ASW, AW
Arbeidsomstandighedenwet	ARBO, ARBW87
Auteurswet 1912	AUTW, AUTEURSW
Bestrijdingsmiddelenwet 1962	BMW62, BMW
Destructiewet	DESW, DSW
Gezondheids- en welzijnswet voor dieren	GWD, GWWD
Hinderwet	HIW, HINDERW, HW
Lozingenbesluit bodembescherming	LBBB, LBB
Monumentenwet 1988	MMW88, MONUMWT
Ontgrondingenwet	ONTW, OGW
Opiumwet	OW, OPIUMW
Scheepvaartverkeerswet	SVVW, SCHVERKW
Telegraaf- en Telefoonwet 1904	TTW04, TELEFW
Toeslagenwet	TW, TOESLW
Vuurwerkbesluit	VWB, VW
Wegenverkeerswet 1994	WVW94, WV
Wet bedreigde uitheemse dier- en plantensoorten	WBUDP, WBD
Wet milieubeheer	WMBH, WM
Wet milieugevaarlijke stoffen	WMGS, WMS [HKS]
Wet op de telecommunicatievoorzieningen	WTCV, WTV
Wetboek van Militair Strafrecht	WMS [OMDATA], MSR, WMSR
Wetboek van Strafrecht	SR, SRTITEL

Tabel Bijlage D.1 – Wetboeken met verschillende afkortingen

Opmerkelijk is dat één afkorting in elke informatiebron gebruikt voor een ander wetboek: de afkorting WMS staat in HKS voor “Wet milieugevaarlijke stoffen” en in OMDATA voor “Wetboek van Militair Strafrecht”.

Verder wordt in HKS de verzamelnaam SRTITEL gebruikt. Een titel is een deel van een wetboek. Elke titel beschrijft een bepaald soort delicten. Om de volgende redenen is het Wetboek van Strafrecht in de vergelijking van wetsartikelen opgedeeld in titels:

- door de omvang is alleen de aanduiding van het wetboek minder sterk dan bij andere wetboeken;
- titels groeperen delicten die bij elkaar horen;
- door deze indeling kan de verzamelnaam SRTITEL geprojecteerd worden op de juiste groep wetsartikelen.

Wetsartikelen

Nu de wetboeken overeenkomen, kunnen de wetsartikelen echt op wetsartikelniveau gereconcilieerd worden. Een wetsartikel bestaat uit een artikelnummer en lidnummers. Deze tweedeling kan voor elk wetsartikel gemaakt worden. Het artikelnummer bestaat namelijk uit het deel van het wetsartikel tot een niet-alfanumeriek teken. Enkele voorbeelden:

Wetsartikel	Wetboek	Artikelnummer	Lidnummers
SR 250A/1/1°	SR 2.14	250a	1/1°
SR 250TER/1/1	SR 2.14	250ter	1/1
ARBTBV 2.4:2/1	ARBTBV	2	4:2/1

Tabel Bijlage D.2 – Splitsing van wetsartikel in artikelnummer en lidnummers

Het onderscheid tussen artikelnummers kan belangrijk zijn. In de opiumwet betekent dit het verschil tussen harddrugs (art. 2) en softdrugs (art. 3).

De lidnummers zijn van het minste belang. Ze voegen weinig nieuwe informatie toe aan de betekenis van het wetsartikel. Bovendien verschilt de notatie per informatiebron en zelfs per wetboek, zoals in de tabel hierboven ook te zien is. Als de lidnummers overeenkomen, dan betekent dit wel meer gelijkheid. Het aandeel zal echter beperkt zijn in de berekening.

Het eerste nummer uit de lidnummers wordt over het algemeen wel consequent ingevoerd en is bovendien met grote zekerheid af te leiden (op een vergelijkbare manier als het artikelnummer). De lidnummers worden daarom gesplitst in een lid en een sublid:

Wetsartikel	Wetboek	Artikelnummer	Lid	Sublid
SR 250A/1/1°	SR 2.14	250a	1	1°
SR 250TER/1/1	SR 2.14	250ter	1	1
ARBTBV 2.4:2/1	ARBTBV	2	4	2/1

Tabel Bijlage D.3 – Splitsing van lidnummers in lid en sublid

Vanwege de verschillende representaties per informatiebron en per wetboek is deze laatste opsplitsing handmatig uitgevoerd. Hierbij is de representatie van de lidnummers consequent doorgevoerd per wetboek, zodat de representatie brononafhankelijk is:

Wetsartikel HKS	Wetsartikel OMDATA	Wetboek	Artikel	Lid	Sublid
SR 250A.1.1	SR 250A/1/1°	SR 2.14	250a	1	1
WM10.44E	WMBH 10.44E	WMBH	10	44e	
WOK30T.1.B	WOK 30T/1/B	WOK	30t	1	b

Tabel Bijlage D.4 – Brononafhankelijke representatie wetsartikelen

Oneigenlijk gebruik

Het is bekend dat de velden voor wetsartikelen in HKS soms oneigenlijk gebruikt worden. Bij oneigenlijk gebruik wordt er andere informatie ingevuld dan wetsartikelen, meestal omdat hier op een andere plaats geen mogelijkheid toe is. Deze informatie is weggefilterd:

Oneigenlijk wetsartikel
AFGPL*
IBBPL*
VOPL*
VOD13
VOGPUDEN
VORHKD

Tabel Bijlage D.5 – Oneigenlijk gebruik van wetsartikelen

Op de plaats van het sterretje komt een viercijferig nummer.

D.2.2 Gelijkheid

De gelijkheid tussen twee wetsartikelen wordt bepaald aan de hand van de opbouw en categorisering van de wetsartikelen.

Categorisering

In de vorige paragraaf is de opbouw van een wetsartikel beschreven in vier onderdelen: wetboek, artikel, lid en sublid. Deze opbouw zegt impliciet iets over het soort delict, omdat wetboeken over bepaalde delicten gaan. Voor een echte categorisering van de wetsartikelen is de opbouw niet bruikbaar:

- sommige categorieën beslaan meerdere wetboeken
- sommige categorieën beslaan meerdere artikelen uit een wetboek

Voor het categoriseren van wetsartikelen worden binnen het WODC twee indelingen gebruikt: de CBS-standaardclassificatie en een eigen indeling van het WODC. Hier is de CBS-standaard-classificatie gebruikt. Deze bestaat uit de volgende categorieën:

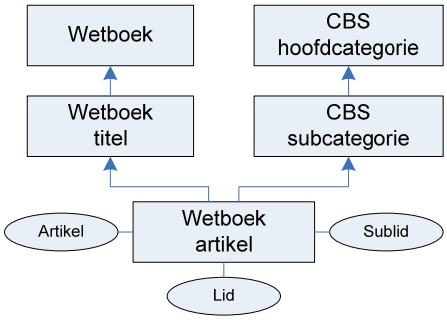
Hoofdcategorie	Subcategorie
Wetboek van Strafrecht	Vernieling en openbare orde Geweldsmisdrijven Vermogensmisdrijven Overige misdrijven
Economische delicten	-
Wegenverkeerswet	Rijden onder invloed Doorrijden na ongeval Overig
Opiumwet	Harddrugs Softdrugs Overig
Wet Wapens en Munitie	-
Wetboek van Militair Strafrecht	-
Overige delicten	-

Tabel Bijlage D.6 – CBS-standaardclassificatie

Behalve een delict kan een wetsartikel ook een algemene bepaling weergeven, zoals ‘poging tot’. Deze wetsartikelen vormen ook een categorie (Wetboek van Strafrecht – Algemene bepalingen), die niet in de CBS-standaardclassificatie voor misdrijven voorkomt.

Berekening

Bij de berekening van een gelijkheidswaarde tussen twee wetsartikelen moet in oogschouw worden genomen dat de wetsartikelen gebruikt worden om een delict te beschrijven. In de loop van de strafrechtsketen kan – bijvoorbeeld door nieuwe bewijzen – de aard van een delict veranderen, waardoor een ander wetsartikel toepasselijker wordt.



Figuur Bijlage D.1 – Onderdelen van een wetsartikel

De gelijkheid van twee wetsartikelen wordt bepaald aan de hand van de eerder genoemde onderdelen, schematisch weergegeven in Figuur Bijlage D.1. Elk onderdeel heeft een bepaald gewicht in de vergelijking:

Onderdeel	Gewicht
CBS standaardclassificatie (hoofdcategorie)	15%
CBS standaardclassificatie (subcategorie)	20%
Wetboek (incl. titel)[9]	35%
Artikel	15%
Lid	10%
Sublid	5%

Tabel Bijlage D.7 – Onderdelen van wetsartikel met gewicht

Het beste gewicht voor de categorieën en het wetboek hangt af van de frequentie waarmee delicten van aard veranderen. Er is hier sprake van een afstandsverdeling: als delicten niet van aard veranderen, dan blijven de artikelen – en daarmee de categorieën en het wetboek – gelijk. De afstand is nul en de frequentie is maximaal. Als de aard van delicten wel verandert, dan houden de meeste delicten binnen hetzelfde wetboek en dezelfde categorieën. De verhouding van het aantal delicten dat van aard verandert ten opzichte van het aantal dat niet van aard verandert, is hier gesteld op 70:85. Mogelijk is deze verhouding kleiner en moet het gewicht van Artikel worden verhoogd ten koste van de gewichten van de categorieën en het wetboek.

Het is bekend onder de bronexperts dat delicten van aard kunnen veranderen. Er is echter geen onderzoek gedaan naar de mate waarin dit gebeurt. De verhoudingen die in dit onderzoek gebruikt zijn, zijn dan ook pure schattingen. Met nader onderzoek kunnen deze verhoudingen beter geschat worden, al blijft het altijd een schatting.

[9] Wetboek en titel zijn als één onderdeel opgenomen. Als er titels binnen een wetboek zijn, dan is de overlap van het wetboek alleen namelijk te zwak.

Bijlage E Implementatie

E.1 Implementatie EROS

E.1.1 Model

Virtuele relaties

Virtuele relaties worden gebruikt om gelijkheidswaarden of gemaakte reconciliaties op te halen uit een eerder geconcilieerde hiërarchische model. Normaal wordt een gelijkheidswaarde berekend in een kennisregel. Bij een virtuele relatie wordt het berekenen van de gelijkheidswaarde vervangen door het ophalen hiervan. In feite is een virtuele relatie dus een speciale variant van een kennisregel, waarin de berekening het ophalen van een gelijkheidswaarde is. Een voorbeeld hiervan staat in paragraaf E.1.3.

Meerwaardige attributen

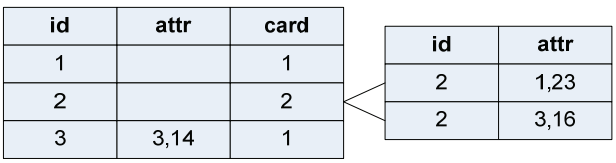
In paragraaf 5.3.2 is gesproken over meerwaardige attributen. Zodra een attribuut mogelijk meerwaardig is in een informatiebron, dan betekent dit in het ontwerp een niveauverschil. In de implementatie is dit efficiënter opgelost.

Stel, een attribuut is gedefinieerd als een meerwaardig attribuut. Mogelijke waarden voor het attribuut zouden kunnen zijn (inclusief percentage van voorkomen), weergegeven als sets:

Waarde	Percentage
{}	10%
{1}	89%
overig	1%

Tabel Bijlage E.1

Voor 99% van de attribuutwaarden is het niet nodig om een niveauverschil te hebben met de entiteit. Door dit gedeelte van de attribuutwaarden wel op te slaan op het entiteitniveau, wordt veel snelheidswinst geboekt. De fysieke implementatie is als volgt:



Figuur Bijlage E.1

Hierbij is *id* de entiteitidentificatie en bevat *attr* de attribuutwaarde. Het attribuut *card* geeft aan of er een niveauverschil is (waarde 2) of niet (waarde 1).

Deze manier van opslag heeft consequenties voor de berekening van het initiële gewicht. Het attribuut bevindt zich immers niet meer in één view, maar in zowel de knoopview als de detailview. Hierdoor moeten de parameters van de id-formule in dit geval berekend worden door de attribuutwaarden uit beide views samen te nemen.

E.1.2 Applicatie

In de applicatie worden een aantal stappen doorlopen, zoals in de architectuur beschreven. Elke stap levert informatie op die van belang is voor het genereren van de scripts.

Views, bronnen en centrumknoop

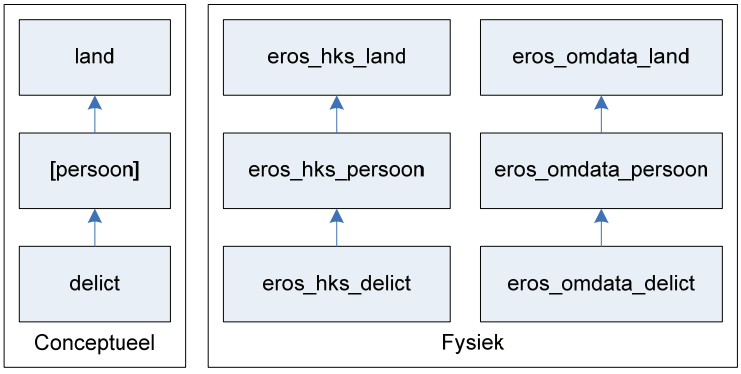
De views worden verondersteld op een bepaalde manier opgebouwd te zijn (zie ontwerp, paragraaf 5.3.1). Hierdoor kan – na selectie van de views – bepaalde informatie worden afgeleid uit de viewdefinities.

Voorbeeld. Stel, de volgende views worden geselecteerd:

View
eros_hks_land
eros_hks_persoon
eros_hks_delict
eros_omdata_land
eros_omdata_persoon
eros_omdata_delict

Tabel Bijlage E.2

Hieruit zijn de volgende conceptuele en fysieke modellen af te leiden:



Figuur Bijlage E.2

Nadat de bronnen zijn geselecteerd (als het voorbeeld HKS als bron I en OMDATA als bron II) en de centrumknoop is gekozen (persoon), kan de volgende informatie worden afgeleid:

Informatie	Templatevariabele	Voorbeeld waarde
Centrumknoop	central_node	persoon
Hoogste knoop	top_node	land
Huidige knoop	node	land (bij knoop Land)
Ouderknoop	parent_node	land (bij knoop Persoon)
Kindknoop	child_node	delict (bij knoop Persoon)
Lijst van knopen	node_list	land, persoon, delict
Bron I	source_1	HKS
Bron II	source_2	OMDATA
View I	view_1	eros_hks_land (bij knoop Land)
View II	view_2	eros_omdata_land (bij knoop Land)

Tabel Bijlage E.3

Verder worden nu de volgende blokvariabelen ondersteund:

Blok	Betekenis
not_top	Dit blok wordt alleen weergegeven als de huidige knoop niet de hoogste knoop is.
low_children	Dit blok wordt herhaald voor elke kindknoop onder de huidige knoop, die meetelt in de gelijkheid van de huidige knoop (de gelijkheid komt van beneden). Binnen dit blok gelden de variabelen: - node (naam van de huidige knoop) - child_node (naam van de kindknoop) - child_number (zie hieronder)
high_child	Dit blok wordt weergegeven als de centrumknoop zich onder de huidige knoop bevindt en de gelijkheid dus naar beneden moet worden doorgegeven. Binnen dit blok gelden de variabelen: - node (naam van de huidige knoop) - child_node (naam van de kindknoop)

Tabel Bijlage E.4

In het voorbeeld hebben deze blokvariabelen de volgende waarde:

Blok	Knoop	Waarde
not_top	land	Niet weergeven
	persoon	Weergeven
	delict	Weergeven
low_children	land	n.v.t.
	persoon	Delict
	delict	n.v.t.
high_child	land	Persoon
	persoon	n.v.t. (niet weergeven)
	delict	n.v.t. (niet weergeven)

Tabel Bijlage E.5

Kennisregels en gelijkheidsfuncties

Selecteren kennisregels
Als laatste stap in de applicatie worden de kennisregels uit het Expert Kennis Systeem geselecteerd. Bij een kennisregel zijn een aantal eigenschappen opgeslagen, welke terugkomen in templatevariabelen:

Templatevariabele	Omschrijving
rule_source_1	Bron I van de kennisregel
rule_source_2	Bron II van de kennisregel
detail_view_1	Detailview in bron I, indien van toepassing.
detail_view_2	Detailview in bron II, indien van toepassing.
merge_function	Samenvoegfunctie (standaard: merge_reconciliate).
merge_use_threshold	Geeft aan of gelijkheid onder de drempel (zie instellingen) moet worden meegenomen in de gelijkheidsmatrix (standaard: nee).
rule_declaration	Naam van de regel zonder prefix; heeft speciale betekenis in mixed-loop functie.

count_1	Telattribuut in bron I (standaard: EROS_COUNT)
count_2	Telattribuut in bron I (standaard: EROS_COUNT)
cardinality_1	Cardinaliteit ten opzichte van de knoop in bron I
cardinality_2	Cardinaliteit ten opzichte van de knoop in bron I
mixed_cp	Optioneel blok; wordt getoond bij meer-op-meer cardinaliteit
mixed_pc1	Optioneel blok; wordt getoond bij 1-op-meer cardinaliteit
mixed_pc2	Optioneel blok; wordt getoond bij meer-op-1 cardinaliteit

Tabel Bijlage E.6

De bronnen van de kennisregel worden opgegeven in het Expert Kennis Systeem en kunnen in volgorde afwijken van de bronnen die later door de gebruiker worden geselecteerd. Als de volgorde gelijk is, dan zijn de variabelen rule_source_1 en rule_source_2 gelijk aan respectievelijk source_1 en source_2; is de volgorde omgedraaid, dan zijn de variabelen gelijk aan respectievelijk source_2 en source_1.

Voor het telattribuut geldt: als het niet is opgegeven, dan wordt gezocht naar het attribuut EROS_COUNT. Als het attribuut niet gevonden wordt, dan krijgt de parameter de waarde 1.

De cardinaliteit geeft aan wat de relatie van de kennisregel met een bron is. Hierbij staat 1 voor een één-op-één relatie en 2 voor een één-op-meer relatie. De optionele mixed-loop functies worden gebruikt als de cardinaliteit wisselt per entiteit (dit komt vooral van pas bij inconsistenties, zie ook paragraaf E.1.1).

Gelijkheidsfuncties in knopen
Nu de kennisregels geselecteerd zijn, kan het aantal gelijkheidsfuncties in elke knoop bepaald worden. Dit aantal wordt namelijk bepaald door het aantal kennisregels in de knoop en het aantal kindknopen dat de gelijkheid naar de knoop doorgeeft.

Templatevariabele	**Omschrijving**
function_count	Totaal aantal gelijkheidsfuncties
function_weight	Lijst met initiële gewichten voor elke functie
rule_count	Aantal kennisregels
knowledge_rules	Blok dat herhaald wordt voor elke kennisregel
rule_number	Nummer van de kennisregel
rule_prefix	Prefix van de kennisregel (voor loop functies)
rule_name	Naam van de kennisregel (zonder prefix)

Tabel Bijlage E.7

Instellingen

Templatevariabele	**Omschrijving**
similarity_factor	Gelijkheidsfactor (zie §3.4)
threshold	Gelijkheid onder deze drempel wordt als ongelijk behandeld
include	Code-include (inhoud is de bestandsnaam)
function	Functie-include (inhoud is de bestandsnaam), de declaratie wordt in het declarations-blok geplaatst

declarations	Blok met functiedeclaraties (voor functies die eerder worden aangeroepen dan dat ze gedeclareerd zijn)
impl_nodes	Blok met code van de knopen (automatisch gegenereerd)
impl_rules	Blok met code van de kennisregels (automatisch gegenereerd)

Tabel Bijlage E.8

E.1.3 Scripts

Het materialisatieplan is een opeenvolging van SQL statements en bevat geen bijzonderheden. Het executieplan wordt samengesteld uit templates, waarbij de templatevariabelen gevuld worden door de applicatie.

In deze paragraaf wordt het voorbeeld uit de vorige paragraaf ook gebruikt.

Opbouw

Het executieplan is een PL/SQL package en bestaat uit een openbare header en een afgeschermde body. In de header staat één publieke procedure ExecutePlan. De package kan als volgt uitgevoerd worden:

```
BEGIN
       EROS_PERSOON.ExecutePlan;
       COMMIT;
END;
```

Deze procedure bestaat uit twee onderdelen: de initialisatie en het eigenlijke reconciliatie-proces. De implementatie wordt uitgelegd met slechts één hiërarchisch model, waardoor er één hoogste knoop wordt aangeroepen:

```
PROCEDURE ExecutePlan IS
BEGIN
        Initialize;
        Node_PERSOON;
END ExecutePlan;
```

Een implementatie met meerdere hiërarchische modellen (en virtuele relaties; een hiërarchisch database model) roept meerdere hoogste knopen aan in een specifieke volgorde. Bijvoorbeeld:

```
PROCEDURE ExecutePlan IS
BEGIN
        Initialize;
        Node_LAND;
        Node_PERSOON;
END ExecutePlan;
```

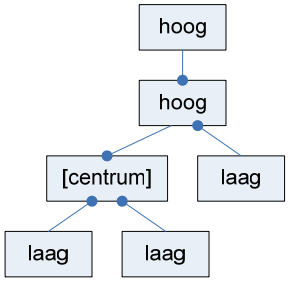
Doordat de enige communicatie tussen de modellen via virtuele relaties geschiedt en de aanroepvolgorde zo gekozen is dat de informatie op het juiste moment aanwezig is, zijn de berekeningen onafhankelijk. Bovenstaande implementatie geeft daarom hetzelfde resultaat als de twee implementaties met één hiërarchisch model (eerst Node_LAND en daarna Node_PERSOON). Voor de uitleg voegt een implementatie met meerdere hiërarchische modellen niets toe.

Knopen

In het hiërarchisch model zijn drie verschillende knopen te onderscheiden:

- de centrumknoop, waar alle gelijkheid samenkomt;
- hoge knopen, die gelijkheid naar beneden doorgeven;
- lage knopen, die gelijkheid naar boven doorgeven.

Een abstract voorbeeld:



Figuur Bijlage E.3

Het belangrijkste verschil tussen hoge en lage knopen is, dat lage knopen een gelijkheidswaarde door moeten geven naar boven. Een centrumknoop kan daarom ook onder de hoge knopen worden gerekend.

De knoopfuncties bevatten – ongeacht het soort knoop – allemaal een carthesisch product van de entiteiten uit beide bronnen:

```
OPEN cur1 FOR
       'SELECT * FROM [EROS::VIEW_1] '||
       'WHERE FK_[EROS::PARENT_NODE]=:pk'
USING PK1;
LOOP
       FETCH cur1 INTO row1;
       EXIT WHEN cur1%NOTFOUND;

       OPEN cur2 FOR
              'SELECT * FROM [EROS::VIEW_2] '||
              'WHERE FK_[EROS::PARENT_NODE]=:pk'
       USING PK2;
       LOOP
              FETCH cur2 INTO row2;
              EXIT WHEN cur2%NOTFOUND;

              ...

       END LOOP;
END LOOP;
```

Door gebruik van het not_top blok rond de where-clausule wordt voorkomen dat deze clausule in de hoogste knoop gebruikt wordt (zie paragraaf E.1.2). Lage knopen bevatten daarnaast code om een gelijkheidswaarde op te leveren.

Binnen het carthesisch product wordt de reconciliatiemogelijkheid berekend tussen twee entiteiten. Allereerst wordt de gelijkheidswaarde geïnitialiseerd, bij hoge knopen (behalve de top) met de gelijkheid afkomstig van de ouder:

Lage knoop of hoogste knoop	Overige hoge knopen

`sim := 0;`
`simcount := 0;`
`simweight := 0;`
	`sim := parentsim;`
	`simcount := 1;`
	`simweight := 1;`

Tabel Bijlage E.9

Vervolgens worden de gelijkheidswaarden van kennisregels berekend. In onderstaande code worden onder andere de volgende variabelen gebruikt:

Variabele	Betekenis
simrule	Array met berekende gelijkheden
simcount	Aantal meegenomen gelijkheden
simweight	Het gewicht van de meegenomen gelijkheden (som)
weights	Array met initiële gewichten

Tabel Bijlage E.10

```
FOR i IN 1..[EROS::RULE_COUNT] LOOP
        simrule(i) := NULL;
        IF (simcount=0) OR (sim > 0) THEN
[EROS::KNOWLEDGE_RULES]
                IF i = [EROS::RULE_NUMBER] THEN
                        simrule(i) :=
                             [EROS::RULE_PREFIX][EROS::RULE_NAME](row[EROS::RULE_SOURCE_1
                             ], row[EROS::RULE_SOURCE_2]);
                END IF;
[/EROS::KNOWLEDGE_RULES]
                IF NVL(simrule(i), -1) = 0 THEN
                        sim := 0;
                        simcount := simcount + 1;
                        simweight := 1;
                ELSIF simrule(i) IS NOT NULL THEN
                        sim := sim + simrule(i)*weights(i);
                        simcount := simcount + 1;
                        simweight := simweight + weights(i);
                END IF;
        END IF;
END LOOP;
```

Te zien is dat kennisregels niet worden meegeteld als ze *n.a.* (NULL) opleveren. Verder wordt het gewicht overgenomen uit de initiële gewichten. Op vergelijkbare manier wordt de gelijkheidswaarde van lage knopen meegenomen:

```
FOR i IN [EROS::RULE_COUNT]+1..[EROS::FUNCTION_COUNT] LOOP
        simrule(i) := NULL;
        IF (simcount=0) OR (sim/simweight > [EROS::THRESHOLD]) THEN
[EROS::LOW_CHILDREN]
```

```
                IF i = [EROS::RULE_COUNT]+[EROS::CHILD_NUMBER] THEN
                        simrule(i) := Node_[EROS::CHILD_NODE](row1.PK_[EROS::NODE],
                            row2.PK_[EROS::NODE]);
                END IF;
[/EROS::LOW_CHILDREN]
                IF NVL(simrule(i), -1) = 0 THEN
                        sim := 0;
                        simcount := simcount + 1;
                ELSIF simrule(i) IS NOT NULL THEN
                        sim := sim + simrule(i)*weights(i);
                        simcount := simcount + 1;
                        simweight := simweight + weights(i);
                END IF;
        END IF;
END LOOP;
```

De gelijkheid wordt uiteindelijk berekend door de totale gewogen gelijkheid te delen door het totale gewicht:

```
IF (simcount > 0) AND (sim > 0) THEN
        sim := sim / simweight;

        ... -- opslag gelijkheid
END IF;
```

Indien de knoop een hoge knoop is, maar geen centrumknoop, dan moet de gelijkheid worden doorgegeven:

```
[EROS::HIGH_CHILD]
Node_[EROS::CHILD_NODE](sim, row1.PK_[EROS::NODE], row2.PK_[EROS::NODE]);
[/EROS::HIGH_CHILD]
```

De gelijkheid wordt opgeslagen in de statistiektabel als deze boven de opslagdrempel uitkomt:

```
IF (SIM > [EROS::THRESHOLD]) THEN
        EXECUTE IMMEDIATE 'INSERT INTO ' || PREFIXS || '[EROS::NODE](' ||
                'PK_[EROS::SOURCE_1], PK_[EROS::SOURCE_2], SIMILARITY' ||
                ') VALUES(' ||
                ROW1.PK_[EROS::NODE] || ', ' || ROW2.PK_[EROS::NODE] || ', :SIM' ||
                ')' USING SIM;
END IF;
```

Op dezelfde manier kan de gelijkheid van een kennisregel worden opgeslagen. In de PL/SQL code wordt de gelijkheid van elke kennisregel opgeslagen in de array *simrule*, op de positie *simrule*(*rule_number*). In onderstaand voorbeeld is *i* het nummer van de kennisregel:

```
IF (simrule(i) > [EROS::THRESHOLD]) THEN
        EXECUTE IMMEDIATE 'INSERT INTO ' || PREFIXR || '[EROS::NODE](' ||
                'PK_[EROS::SOURCE_1], PK_[EROS::SOURCE_2], RULE, SIMILARITY' ||
```

```
                ') VALUES(' ||
                ROW1.PK_[EROS::NODE] || ',' || ROW2.PK_[EROS::NODE] || ',:RULE,:SIM' ||
                ')' USING i, simrule(i);
END IF;
```

Niveauverschillen

In de knopen worden kennisregels aangeroepen door:

```
[EROS::RULE_PREFIX][EROS::RULE_NAME]
```

De prefix is afhankelijk van het niveauverschil van de kennisregel met de entiteit:

Prefix	Niveau in bron I	Niveau in bron II
RULE_	1	1
LOOP_CP_	M	N
LOOP_PC1_	1	N
LOOP_PC2_	M	1
MIXED_LOOP_	Wisselend	Wisselend

Tabel Bijlage E.11

Mixed-loop functie

De mixed-loop functie kijkt voor elke entiteitcombinatie welk niveauverschil er is:

```
FUNCTION LOOP_MIXED_[EROS::RULE_NAME](ROW1 [EROS::VIEW_1]%ROWTYPE, ROW2
                    [EROS::VIEW_2]%ROWTYPE) RETURN NUMBER
IS
-- mixed merge function
       card1 INTEGER;
       card2 INTEGER;
       sim NUMBER;
BEGIN
       card1 := [EROS::CARDINALITY_1];
       card2 := [EROS::CARDINALITY_2];

       CASE
               [EROS::MIXED_CP]WHEN card1 = 2 AND card2 = 2 THEN sim :=
                      LOOP_CP_[EROS::RULE_NAME](ROW1, ROW2);[/EROS::MIXED_CP]
               [EROS::MIXED_PC1]WHEN card1 = 1 AND card2 = 2 THEN sim :=
                      LOOP_PC1_[EROS::RULE_NAME](ROW1, ROW2);[/EROS::MIXED_PC1]
               [EROS::MIXED_PC2]WHEN card1 = 2 AND card2 = 1 THEN sim :=
                      LOOP_PC2_[EROS::RULE_NAME](ROW1, ROW2);[/EROS::MIXED_PC2]
       ELSE
              sim := RULE_[EROS::RULE_NAME](ROW1, ROW2);
       END CASE;

       RETURN sim;
END LOOP_MIXED_[EROS::RULE_NAME];
```

Een loopfunctie kan alleen gedefinieerd worden op een detailview. Stel de kennisregel heeft geen detailview in bron I, omdat de cardinaliteit altijd 1:1 is. Dan zijn de loopfuncties LOOP_CP en LOOP_PC2 niet mogelijk. De blokken MIXED_CP, MIXED_PC1 en MIXED_PC2 zorgen ervoor dat de juiste loopfuncties gedeclareerd worden.

Overige loopfuncties
De loopfuncties CP, PC1 en PC2 voeren de kennisregel uit voor elke attribuutwaarde. Vervolgens worden de opgeleverde gelijkheidswaarden samengevoegd tot één gelijkheidswaarde.

Kennisregel declaratie
Uiteindelijk wordt de expertkennis geïmplementeerd in de functie `RULE_[EROS::RULE_NAME]`. Kennisregels waarbij het niveauverschil gewoon voor alle entiteiten gelijk is, worden gedeclareerd met *rowtypes* als parameters. Voor een kennisregel Land bijvoorbeeld:

```
FUNCTION RULE_land(row1 EROSM_HKS_LAND%ROWTYPE, row2 EROSM_OMDATA_LAND%ROWTYPE) RETURN
    NUMBER IS
BEGIN
       RETURN CASE WHEN row1.NAAM = row2.Naam THEN 1 ELSE 0 END;
END RULE_land;
```

Voor kennisregels waarbij het niveauverschil per entiteit kan verschillen, moeten alle varianten gedeclareerd worden. Elke variant verwijst vervolgens naar de eigenlijke implementatie:

```
FUNCTION RULE_geslacht(val1 VARCHAR2, val2 VARCHAR2) RETURN NUMBER IS
BEGIN
       RETURN CASE
               WHEN val1 IS NULL OR val2 IS NULL THEN NULL
               WHEN val1 = val2 THEN 1
       ELSE
               0
       END;
END RULE_geslacht;

FUNCTION RULE_CP_geslacht(row1 EROSM_HKS_PERSOON_GESLACHT%ROWTYPE, row2
    EROSM_OMDATA_PERSOON_GESL%ROWTYPE) RETURN NUMBER IS
BEGIN
       RETURN RULE_geslacht(row1.PK_GESLACHT, row2.PK_GESLACHT);
END RULE_CP_geslacht;

FUNCTION RULE_PC1_geslacht(row1 EROSM_HKS_PERSOON%ROWTYPE, row2
    EROSM_OMDATA_PERSOON_GESL%ROWTYPE) RETURN NUMBER IS
BEGIN
       RETURN RULE_geslacht(row1.GESLACHT, row2.PK_GESLACHT);
END RULE_PC1_geslacht;

FUNCTION RULE_PC2_geslacht(row1 EROSM_HKS_PERSOON_GESLACHT%ROWTYPE, row2
    EROSM_OMDATA_PERSOON%ROWTYPE) RETURN NUMBER IS
BEGIN
       RETURN RULE_geslacht(row1.PK_GESLACHT, row2.GESLACHT);
```

```
END RULE_PC2_geslacht;

FUNCTION RULE_geslacht(row1 EROSM_HKS_PERSOON%ROWTYPE, row2
     EROSM_OMDATA_PERSOON%ROWTYPE) RETURN NUMBER IS
BEGIN
       RETURN RULE_geslacht(row1.GESLACHT, row2.GESLACHT);
END RULE_geslacht;
```

Bovenstaande kennisregel is een voorbeeld van inconsistentie, in dit geval in het attribuut Geslacht. Merk op dat voor de variant zonder niveauverschil nog steeds de gewone declaratie (RULE_geslacht) kan worden gebruikt.

Kennisregels

In de vorige paragraaf is de implementatie van een kennisregel uitgewerkt tot het deel dat de expertkennis bevat. Deze paragraaf laat zien hoe verschillende soorten expertkennis geïmplementeerd kunnen worden in een kennisregel.

Voor alle kennisregels geldt: als één van de waarden NULL (*n.a.*) is, dan levert de kennisregel ook NULL op. De eenvoudigste implementatie van een kennisregel is de exacte vergelijking:

```
FUNCTION RULE_geslacht(val1 VARCHAR2, val2 VARCHAR2) RETURN NUMBER IS
BEGIN
       RETURN CASE
              WHEN val1 IS NULL OR val2 IS NULL THEN NULL
              WHEN val1 = val2 THEN 1
       ELSE
              0
       END;
END RULE_geslacht;
```

Een afstandsverdeling rond een verwachte afstand 0 kan als volgt geïmplementeerd worden:

```
FUNCTION RULE_verdeling(val1 DATE, val2 DATE) RETURN NUMBER IS
       diff INTEGER;
       sim NUMBER;
BEGIN
       if (val1 IS NULL) OR (val2 IS NULL) then
              sim := NULL;
       else
              diff := val1-val2;
              sim := exp(-diff*diff/50);
       end if;

       RETURN sim;
END RULE_verdeling;
```

Hetzelfde voorbeeld, alleen nu met een ruisdrempel:

```
FUNCTION RULE_verdeling(val1 DATE, val2 DATE) RETURN NUMBER IS
	diff INTEGER;
	sim NUMBER;
BEGIN
	if (val1 IS NULL) OR (val2 IS NULL) then
		sim := NULL;
	else
		diff := val1-val2;
		sim := exp(-diff*diff/50);
		if (sim <= [EROS::THRESHOLD]) then
			sim := [EROS::THRESHOLD]+0.01;
		end if;
	end if;

	RETURN sim;
END RULE_verdeling;
```

Virtuele relaties worden ook als kennisregel geïmplementeerd. Onderstaande voorbeelden zijn virtuele relaties naar het eerder geconcilieerde Land en leveren respectievelijk een gelijkheidswaarde of een reconciliatiewaarde (0 of 1) op.

```
FUNCTION RULE_virtueel_sim(code1 INTEGER, code2 INTEGER) RETURN NUMBER IS
	sim NUMBER;
	TYPE ProbCurTyp IS REF CURSOR;
	prob_cv ProbCurTyp;
BEGIN
	IF (code1 IS NULL) OR (code2 IS NULL) THEN
		sim := NULL;
	ELSE
		OPEN prob_cv FOR
			'SELECT SIMILARITY '||
			'FROM EROSS_LAND ' ||
			'WHERE PK_BRON1=:pk1 AND PK_BRON2=:pk2'
		USING code1, code2;
		FETCH prob_cv INTO sim;
		IF prob_cv%NOTFOUND THEN
			sim := 0;
		END IF;
	END IF;

	RETURN sim;
END RULE_virtueel_sim;

FUNCTION RULE_virtueel_recon(code1 INTEGER, code2 INTEGER) RETURN NUMBER IS
	sim NUMBER;
	TYPE ProbCurTyp IS REF CURSOR;
	prob_cv ProbCurTyp;
BEGIN
	IF (code1 IS NULL) OR (code2 IS NULL) THEN
```

```
                sim := NULL;
        ELSE
                OPEN prob_cv FOR
                        'SELECT RECONCILIATED '||
                        'FROM EROSS_LAND ' ||
                        'WHERE PK_BRON1=:pk1 AND PK_BRON2=:pk2'
                USING code1, code2;
                FETCH prob_cv INTO sim;
                IF prob_cv%NOTFOUND THEN
                        sim := 0;
                END IF;
        END IF;

        RETURN sim;
END RULE_virtueel_recon;
```

Tenslotte een voorbeeld met meerdere afstandsverdelingen. Hierbij is voor snelrechtzaken een identificerende sleutel beschikbaar, terwijl overige zaken worden vergeleken met een datumvergelijking:

```
FUNCTION RULE_verdelingen(row1 EROSM_HKS_ZAAK%ROWTYPE, row2 EROSM_OMDATA_ZAAK%ROWTYPE)
RETURN NUMBER IS
        diff INTEGER;
        sim NUMBER;
BEGIN
        if (row1.IS_SNELRECHT IS NULL) OR (row2.IS_SNELRECHT IS NULL) then
                -- niet beschikbaar
                sim := NULL;
        else
        if (row1.IS_SNELRECHT = 1) and (row2.IS_SNELRECHT = 1) then
                if (row1.SNEL_ID IS NULL) OR (row2.SNEL_ID IS NULL) then
                        -- niet beschikbaar
                        sim := NULL;
                else
                        sim := CASE WHEN row1.SNEL_ID=row2.SNEL_ID THEN 1 ELSE 0 END;
                end if;
        else
        if (row1.IS_SNELRECHT = 0) and (row2.IS_SNELRECHT = 0) then
                if (row1.DATUM IS NULL) OR (row2.DATUM IS NULL) then
                        -- niet beschikbaar
                        sim := NULL;
                else
                        diff := row1.DATUM - row2.DATUM;
                        sim := exp(-diff*diff/50);
                end if;
        else
                -- niet vergelijkbaar
                sim := NULL;
        end if;
```

```
        RETURN sim;
END RULE_verdelingen;
```

E.2 Implementatie casus in EROS

Deze paragraaf beschrijft hoe de casus is uitgewerkt in EROS.

E.2.1 Gemeenschappelijke basis

Zoals eerder vastgesteld in paragraaf 4.2 bestaat het informatiemodel uit drie hiërarchische modellen: *S*(*Land*), *S*(*Persoon*) en *S*(*Wetsartikel*). De toepassing van EROS op de casus richt zich op de reconciliatie van personen in het hiërarchische model *S*(*Persoon*). Voordat dit kan gebeuren, moeten eerst de andere modellen geconcilieerd zijn. Dit valt buiten de casus, maar is wel besproken in Bijlage D.

Aanvankelijk waren de landen puur op basis van naam gereconcilieerd. Tijdens de analyse bleek dat de impact van landen die zijn samengegaan of uit elkaar gevallen, opvallend groot was. Uiteindelijk zijn alle landen zodanig gereconcilieerd dat landen die een bepaalde geschiedenis met elkaar hebben (koloniaal, opsplitsing, etc.) een bepaalde mate van gelijkheid hebben.

Het reconciliëren van de wetsartikelen is volledig gelukt op de manier zoals dit gewenst was, namelijk door wetsartikelen een bepaalde mate van gelijkheid te geven, wanneer ze mogelijk hetzelfde delict beschrijven.

E.2.2 Informatiemodel

De eerste paragraaf beschrijft welke attributen uit de informatiebronnen gebruikt zijn aan de hand van het fysieke datamodel. In de volgende paragraaf zal verder worden ingegaan op het identificeren van entiteiten.

E.2.2.1 Fysiek datamodel

HKS

Uit de informatiebron HKS zijn de volgende tabellen gebruikt:

Tabel	Verkorte naam	Entiteittype-niveau
ANT_DAT_INV	Ant_dat	proces-verbaal
LANDEN0	Landen	land
LAND_ANT_DEL_TOT9603	Ant_del	delict
LAND_DELIKT_TOT9603	Delict	delict
LAND_RESULT_TOT9603	Result	persoon

Tabel Bijlage E.12

Alle tabellen zijn ingedeeld in peiljaren, waardoor gegevens herhaaldelijk en mogelijk inconsistent kunnen voorkomen. De verkorte naam zal in dit document gebruikt worden.

Uit de bovenstaande tabellen worden de volgende gegevens gehaald:

Attribuut	Tabel
ID Persoon	- metanr (versie 2004) en pers_id in Ant_dat - metanr en persid in Ant_del - metanr in Delict en Result
ID Proces-verbaal	Ant_dat, Ant_del, Delict
ID Delict	Ant_del, Delict
Geboortedatum	Result
Geslacht	Result
Geboorteland	Result
Land	Landen
Datum proces-verbaal	Ant_dat
Wetsartikel	Delict

Tabel Bijlage E.13

Het entiteittype Delict staat hier wel genoemd, omdat de wetsartikelen op dit niveau opgeslagen zijn. In het reconciliatieproces speelt dit entiteittype uiteindelijk geen rol.

OMDATA

Uit de informatiebron OMDATA zijn de volgende tabellen gebruikt:

Tabel	Entiteittype
FEIT	delict
ZAAK	zaak

Tabel Bijlage E.14

Uit de bovenstaande tabellen worden de volgende gegevens gehaald:

Attribuut	Tabel
ID Persoon	Via OBJD (zaakniveau)
ID Proces-verbaal	Feit
ID Delict	Feit
Geboortedatum	Zaak
Geslacht	Zaak
Geboorteland	Zaak
Land	Zaak
Pleegdatum	Feit
Wetsartikel	Feit

Tabel Bijlage E.15

Ook hier geldt dat het entiteittype Delict hier wel genoemd wordt vanwege de wetsartikelen, maar dit entiteittype in het reconciliatieproces uiteindelijk geen rol speelt. Daarnaast wordt uit de tabel Feit – horende bij het entiteittype Delict – het entiteittype Proces-verbaal afgeleid.

E.2.2.2 Afleiding entiteittypen

In zowel HKS als OMDATA komen niet alle gemeenschappelijke entiteittypen terug als tabel in het oorspronkelijke fysieke datamodel. Deze paragraaf bespreekt hoe deze entiteittypen afgeleid zijn uit dit oorspronkelijke datamodel en hoe de bijbehorende entiteiten geïdentificeerd zijn. Zie voor de afleiding van landen en wetsartikelen Bijlage D.

HKS

Binnen HKS is het *peiljaar* een belangrijk gegeven. Een peiljaar is het jaar waarop een proces-verbaal is ingevoerd. Processen-verbaal kunnen in meerdere peiljaren worden ingevoerd (zie afleiding entiteittype Proces-verbaal).

Persoon
In HKS kunnen personen op twee manieren onderscheiden worden:

- via een metanummer, dat landelijk uniek is
- via een persoonsnummer, dat regionaal en per peiljaar uniek is

Alle tabellen bevatten een metanummer. Dit nummer is consistent binnen één versie van het HKS. De tabel met delicten (ANT_DEL) bevat het persoonsnummer. Dit nummer is consistent, ook in verschillende versies van HKS.

In dit onderzoek zijn twee verschillende versies van het HKS gebruikt. Aanvankelijk is gestart met een versie uit 2003, omdat de metanummers uit deze versie aansluiten op de metanummers uit de referentieset. Later is er een nieuwe tabel bijgekomen (ANT_DAT) met twee belangrijke attributen: geboortedatum en datum antecedent. Deze tabel is afkomstig uit een nieuwe versie van HKS uit 2004, waarin het metanummer anders is opgebouwd. Deze tabel is gekoppeld aan de hand van de persoonsnummers in ANT_DEL.

Samengevat: personen worden geïdentificeerd aan de hand van het metanummer uit 2003.

Proces-verbaal
Een proces-verbaal is een zwak entiteittype onder persoon en wordt – naast de persoonsidentificatie – geïdentificeerd door het peiljaar en een antecedent-id. Processen-verbaal kunnen echter meerdere keren voorkomen. Elk jaar dat een verdachte een delict pleegt, wordt namelijk de geschiedenis van deze verdachte opnieuw ingevoerd. In de volgende tabel zijn de regels met hetzelfde peiljaar hetzelfde proces-verbaal:

Metanr	Jaar	Peiljaar	Antecedent-id
123	1997	1997	001
123	1999	1997	001
123	1999	1999	001
123	2000	1997	001
123	2000	1999	001
123	2000	2000	001

Tabel Bijlage E.16 – Dubbele processen-verbaal in HKS

In de casus zijn dubbele voorkomens van processen-verbaal verwijderd. Voor elke tabel met dit probleem is een nieuwe tabel aangemaakt met unieke processen-verbaal. In de meeste gevallen zijn dit de records waarvoor het jaar gelijk is aan het peiljaar. Voor sommige processen-verbaal

bestaan deze records echter niet. Voor deze processen-verbaal is het record genomen van het eerst voorkomende jaar.

Samengevat: na ontdubbeling worden de processen-verbaal uniek geïdentificeerd door de persoonsidentificatie, peiljaar en antecedent-id.

OMDATA

Persoon
Personen worden afgeleid uit de OBJD, zoals eerder beschreven in paragraaf 4.3.

Proces-verbaal
OMDATA bevat geen goede mogelijkheid om processen-verbaal te identificeren. Er is gezocht naar een manier om de werkelijkheid zo goed mogelijk te benaderen.

Op feitniveau is een procesverbaal-nummer aanwezig. Dit is geen unieke identificatie. Bovendien wordt het niet altijd (correct) ingevuld. Volgens de definitie van het Openbaar Ministerie kunnen onder een zaak één of meerdere processen-verbaal vallen (zie paragraaf 2.1.2). Proces-verbaal kan daarom gezien worden als een zwak entiteittype onder Zaak, met als zwakke sleutel het procesverbaal-nummer. Processen-verbaal zonder nummer worden samengenomen, omdat zij geen sleutel hebben en onderscheid dus niet mogelijk is.

E.2.2.3 Inconsistentie en missende informatie

Sommige attributen moesten geaggregeerd worden tot één attribuutwaarde. Door inconsistentie in de informatiebronnen kon niet voor elk attribuut een unieke waarde bepaald worden.

Bij de volgende attributen is er sprake van inconsistentie:

Bron	Attribuut	Aantal	Percentage
HKS	Geboortedatum	1 uit 10.000	0‰
	Geslacht	33 uit 10.000	3‰
	Geboorteland	77 uit 10.000	8‰
OMDATA	Geboortedatum	136 uit 8.705	16‰
	Geslacht	61 uit 8.705	7‰
	Geboorteland	121 uit 8.705	14‰
	Geboorteland (na optimalisatie)	136 uit 8.705	16‰

Tabel Bijlage E.17 – Attributen met inconsistentie

Opvallend is dat OMDATA een hogere inconsistentie heeft dan HKS. Bronexperts hadden op basis van hun expertkennis het tegenovergestelde verwacht. Dit komt waarschijnlijk doordat OMDATA geen persoonsentiteiten bevat, waardoor de inconsistentie op persoonsniveau in OMDATA nooit naar boven is gekomen.

Attributen kunnen alleen meegenomen worden in kennisregels als de waarde bekend is. Voor sommige attributen was echter geen waarde beschikbaar.

Bron	Attribuut	Aantal	Promillage
HKS	Geboortedatum	33 uit 10.000	3‰
	Geslacht	0 uit 10.000	0‰

	Geboorteland	42 uit 10.000	4‰
OMDATA	Geboortedatum	16 uit 8.705	2‰
	Geslacht	3 uit 8.705	0‰
	Geboorteland	10 uit 8.705	1‰
	Pleegdatum	2 uit 42.749	0‰

Tabel Bijlage E.18 – Attributen met missende informatie

E.2.3 Kennisregels

In paragraaf 4.1 is de expertkennis besproken die gebruikt wordt in deze casus. De algemene beschrijvingen van kennisregels die daar genoemd zijn, worden nu ingevoerd in het Expert Kennis Systeem. Hiervoor is de specificatie van de kennisregels nodig (de formalisatie van de gelijkheidsberekening) en de koppeling met de informatiebronnen: attribuutnamen en locatie in de bronnen, eventuele samenvoegfuncties, etc. Voor de duidelijkheid worden de specificaties, zoals ze in paragraaf 4.1 afgeleid zijn, nogmaals weergegeven.

Het omzetten van kennisregels naar PL/SQL code gebeurt vanuit het Expert Kennis Systeem, aan de hand van de ingevoerde eigenschappen. Zie E.3 (Implementatie kennisregels) voor meer informatie.

Voor alle kennisregels geldt (tenzij anders vermeld):

Kenmerk	**Waarde**
$bron_1$	HKS
$bron_2$	OMDATA
$aantal_1$	(automatisch)
$aantal_2$	(automatisch)
gewicht	(automatisch)
ruisdrempel	(geen)
samenvoegfunctie	(automatisch)

Tabel Bijlage E.19 – Standaard eigenschappen kennisregels

Attributen worden voor een deel vastgelegd door andere eigenschappen. In de naamgeving wordt daarom alleen het deel genoemd dat een attribuut onderscheidt. Bijvoorbeeld:

Het attribuut Geslacht bevindt zich onder entiteittype Persoon in de bron HKS. Fysiek bevindt dit attribuut zich in HKS_PERSOON.geslacht. Echter, in de kennisregel kan 'HKS_PERSOON' al worden afgeleid van de andere eigenschappen. Daarom wordt het attribuut Geslacht slechts aangeduid met 'geslacht'.

Het attribuut Wetsartikel bevindt zich onder entiteittype PV in de bron OMDATA. Fysiek bevindt het attribuut zich (vanwege de één-op-meer relatie) in OMDATA_PV_WA.pk_wa. Het eerste gedeelte 'OMDATA_PV_' kan afgeleid worden via andere eigenschappen. Het attribuut wordt daarom aangeduid met 'WA.pk_wa'.

Geboortedatum

Specificatie kennisregel:

$$sim = \begin{cases} n.a. & \text{als } datum_1 = n.a. \vee datum_2 = n.a. \\ 0 & \text{als } datum_1 \neq datum_2 \\ 1 & \text{als } datum_1 = datum_2 \end{cases}$$

Na optimalisatie:

$$sim = \begin{cases} n.a. & \text{als } datum_2 = n.a. \vee (datum_1 = n.a. \wedge jaar_1 = n.a.) \\ 0 & \text{als } datum_1 \neq datum_2 \\ 0.5 & \text{als } jaar_1 = jaar(datum_2) \\ 1 & \text{als } datum_1 = datum_2 \end{cases}$$

Kenmerk	Waarde
naam	gebdatum
knoop	persoon
$attr_1$	$datum_1$
$attr_2$	$datum_2$

Tabel Bijlage E.20 – Kennisregel geboortedatum (eigenschappen)

Attribuut	Herkomst
$datum_1$	gebdat of gebdat.pk_gebdat
$jaar_1$	gebjaar
$datum_2$	gebdat of gebdat.pk_gebdat

Tabel Bijlage E.21 – Kennisregel geboortedatum (attributen)

Geslacht

Specificatie kennisregel:

$$sim = \begin{cases} n.a. & \text{als } geslacht_1 = n.a. \vee geslacht_2 = n.a. \\ 0 & \text{als } geslacht_1 \neq geslacht_2 \\ 1 & \text{als } geslacht_1 = geslacht_2 \end{cases}$$

Kenmerk	Waarde
naam	geslacht
knoop	persoon
$attr_1$	$geslacht_1$
$attr_2$	$geslacht_2$

Tabel Bijlage E.22 – Kennisregel geslacht (eigenschappen)

Attribuut	Herkomst
$geslacht_1$	geslacht of geslacht.pk_geslacht
$geslacht_2$	geslacht of geslacht.pk_geslacht

Tabel Bijlage E.23 – Kennisregel geslacht (attributen)

Geboorteland

Specificatie kennisregel:

$$sim = \begin{cases} n.a. & \text{als } gebland_1 = n.a. \vee gebland_2 = n.a. \\ 0 & \text{als } gebland_1 \neq gebland_2 \\ 1 & \text{als } gebland_1 = gebland_2 \end{cases}$$

Na optimalisatie:

$$sim = \begin{cases} n.a. & \text{als } gebland_1 = n.a. \vee gebland_2 = n.a. \\ sim_{land}(gebland_1, gebland_2) & \text{anders} \end{cases}$$

waarbij

$$sim_{land} = \begin{cases} 1 & \text{als } land_1 = land_2 \\ 0.9 & \text{als } land_1 = land_2 \wedge kleingrondgebied_1 \neq kleingrondgebied_2 \\ 0.8 & \text{als } land_1 = land_2 \wedge grootgrondgebied_1 \neq grootgrondgebied_2 \\ 0.5 & \text{als } land_x = \text{V.A.E.} \wedge land_y \in (\text{Egypte}, \text{Syrië}) \\ 0.5 & \text{als } land_x = \text{Duits Oost} - \text{Afrika} \wedge land_y \in (\text{Burundi}, \text{Rwanda}, \text{Tanganyika}) \\ 0.3 & \text{als } land_x = \text{Oost Afrika} \wedge land_y \in (\text{Kenia}, \text{Oeganda}, \text{Tanzania}) \\ 0.3 & \text{als } land_x = \text{Oostenrijk - Hongarije} \wedge land_y \in (\text{Oostenrijk}, \text{Hongarije}) \\ 0 & \text{anders} \end{cases}$$

met: $x \in (1,2) \wedge y \in (1,2) \wedge x \neq y$

Kenmerk	Waarde
naam	gebland
knoop	persoon
$attr_1$	$gebland_1$
$attr_2$	$gebland_2$

Tabel Bijlage E.24

Attribuut	Herkomst
$gebland_1$	gebland of gebland.pk_gebland
$gebland_2$	gebland of gebland.pk_gebland

Tabel Bijlage E.25

Pleegdatum

Specificatie kennisregel:

$$sim = \begin{cases} 0 & \delta < 0 \\ e^{-\delta^2/50} & \delta >= 0 \end{cases}$$

met

$$\delta = datum_{pv} - datum_{delict}$$

Kenmerk	Waarde
naam	antdatum
knoop	pv
$attr_1$	$antdat_1$
$attr_2$	$dtplegen_2$
ruisdrempel	1 jaar (na optimalisatie)

Tabel Bijlage E.26

Attribuut	Herkomst
$antdat_1$	antdat

$dtplegen_2$	dtplegen of dtplegen.pk_dtplegen

Tabel Bijlage E.27

Wetsartikelen

Kenmerk	Waarde
naam	wa
knoop	pv
$attr_1$	wa_1
$attr_2$	wa_2
samenvoegfunctie	merge_wa (na optimalisatie)

Tabel Bijlage E.28

Attribuut	Herkomst
wa_1	wa of wa.pk_wa
wa_2	wa of wa.pk_wa

Tabel Bijlage E.29

De wetsartikelen zijn vooraf gereconcilieerd. De kennisregel voor wetsartikelen in de persoonsreconciliatie is daarom de implementatie van een virtuele relatie. De gelijkheid afkomstig van virtuele relaties kan echter ook geoptimaliseerd worden. Vanwege het niveauverschil is namelijk een samenvoegfunctie nodig. Deze samenvoegfunctie is voor de tweede ronde geoptimaliseerd.

De standaard samenvoegfuncties ($m(K)$/*merge_reconciliate* en $m_p(K)$/*merge_reconciliate_all*) berekenen de gelijkheid aan de hand van gereconcilieerde waarden. Dit impliceert dat elk wetsartikel in bron I hoogstens met één wetsartikel in bron II gereconcilieerd wordt en vice versa. Een eigenschap van wetsartikelen is echter dat ze in de loop van de strafrechtsketen nog wel eens worden gesplitst in wetsartikelen die sterk lijken op het oorspronkelijke wetsartikel. Ook het samenvoegen van wetsartikelen kan voorkomen.

Het is dus niet zozeer belangrijk dat elk wetsartikel een overeenkomstig wetsartikel in de andere bron heeft. Belangrijker is dat elk wetsartikel uit de ene bron gelijkheid heeft met minstens één wetsartikel in de andere bron.

Voor elke bron kan deze nieuwe definitie van gelijkheid eenvoudig worden berekend. Stel de huidige vergelijking heeft in de bron n wetsartikelen. Voor elk wetsartikel wa_i ($i \in \{1..n\}$) gelden de functies:

$simset_i$ de set van gelijkheidswaarden met de andere bron
occ_i het aantal voorkomens van het wetsartikel in de bron

De gelijkheidswaarde sim_{bron} van een bron is nu gedefinieerd als:

Error! Objects cannot be created from editing field codes.

In definitie geldt dat de gelijkheidswaarde maximaal (1) is, als elk wetsartikel de maximale gelijkheidswaarde heeft met minimaal één wetsartikel in de andere bron. Als er wetsartikelen zijn waarbij geen gelijkheid wordt gevonden, dan tellen deze alleen mee in de noemer. Hierdoor is de gelijkheidswaarde relatief ten opzichte van de gelijkheidswaarde die maximaal gehaald zou kunnen worden. Dit is vergelijkbaar met de samenvoegfunctie *merge_reconciliate_all* (of $m_p(K)$, zie Bijlage C).

De uiteindelijke gelijkheid wordt bepaald door het minimum van de gelijkheidswaarden afkomstig van elke bron:

$sim = \min(sim_I, sim_{II})$

E.2.4 Uitvoering

Alle benodigde informatie is verzameld is, kan het reconciliatieproces worden uitgevoerd. Dit is een aantal keer gebeurd. Allereerst is het reconciliatieprogramma EROS getest met behulp van de casus (testfase). Daarna zijn de kennisregels van de casus zelf verbeterd aan de hand van de resultaten uit het reconciliatieproces (optimalisatiefase).

Testfase

De gelijkheidswaarde die de verschillende kennisregels opleveren bij bepaalde vergelijkingen is bekend uit de specificaties. Uiteindelijk vormen de verschillende gelijkheidswaarden de objectgelijkheid waarmee de reconciliaties gekozen worden. Het reconciliatieprogramma is getest door elke stap van kennisregel naar objectgelijkheid één voor één te testen:

- gelijkheidsfunctie van de kennisregel
- gelijkheidswaarde van de kennisregel: samenvoeging van uitkomsten gelijkheids-functies
- entiteitgelijkheid: samenvoeging van gelijkheidswaarden van kennisregels
- objectgelijkheid: samenvoeging van oudergelijkheid, entiteitgelijkheid en kind-gelijkheid

Details worden besproken in de documentbijlage EROS.

Optimalisatiefase

De optimalisatie van kennisregels is al eerder besproken. Het executieplan is echter ook op enkele punten geoptimaliseerd. Optimalisatie van het script was echter geen belangrijk issue in het onderzoek, waardoor hier waarschijnlijk nog veel verbetering mogelijk is.

Het reconciliatieproces bestaat uit vele gelijkheidsberekeningen. Hiervoor is veel tijd nodig. Het eerste proces is niet voltooid, maar zou er waarschijnlijk ongeveer een week over gedaan hebben. Hierna is de optimalisatie voor meerwaardige attributen geïmplementeerd (zie paragraaf E.1.1). Dit verminderde het aantal benodigde queries aanzienlijk. Het uitvoeren van het reconciliatieproces met alleen de queries duurde nu slechts enkele minuten.

Na deze optimalisatie duurde het proces nog twee werkdagen. De oorzaak bleek te liggen in het samenvoegen van gelijkheid. Voor een efficiënte uitvoering zijn index-bepalingen nodig. PL/SQL ondersteunt deze echter niet in standaardfuncties. Het proces slaat de gelijkheids-waarden daarom tijdelijk op in een tabel, waarna met een aantal queries het resultaat berekend kan worden. Waarschijnlijk zou er een aanzienlijke snelheidswinst behaald worden als het samenvoegen van gelijkheidswaarden via een externe procedure in het geheugen zou gebeuren.

Naast het optimaliseren van prestaties is ook de gebruikersvriendelijkheid tijdens de uitvoering verbeterd. Zo is er een script ontwikkeld om de voortgang van het proces in de gaten te houden. Ook kan het proces tussentijds onderbroken worden en eenvoudig weer worden hervat[10].

E.3 Implementatie kennisregels

E.3.1 Kennisregel 'geboortedatum'

```
FUNCTION RULE_gebdatum(date1 DATE, year1 INTEGER, date2 DATE) RETURN NUMBER IS
  sim NUMBER;
BEGIN
  IF date2 IS NULL THEN
    sim := NULL;
  ELSIF date1 IS NULL THEN
    sim := CASE
              WHEN year1 IS NULL THEN NULL
              WHEN year1 = TO_NUMBER(TO_CHAR(date2, 'YYYY')) THEN 0.5
       ELSE
              0
       END;
  ELSE
    sim := CASE
              WHEN date1 IS NULL THEN NULL
              WHEN date1 = date2 THEN 1
       ELSE
              0
       END;
  END IF;
  RETURN sim;
END RULE_gebdatum;
```

E.3.2 Kennisregel 'geslacht'

```
FUNCTION RULE_geslacht(val1 VARCHAR2, val2 VARCHAR2) RETURN NUMBER IS
BEGIN
       RETURN
       CASE
              WHEN val1 IS NULL OR val2 IS NULL THEN NULL
              WHEN val1 = val2 THEN 1
       ELSE
              0
       END;
END RULE_geslacht;
```

E.3.3 Kennisregel 'geboorteland'

De landen zijn grotendeels handmatig gereconcilieerd. Zie voor een uitgebreide beschrijving van dit proces de bijlage 'Reconciliatie landen'. Aanvankelijk had de kennisregel als enige uitkomsten NULL, 0 of 1:

```
FUNCTION RULE_gebland(val1 INTEGER, val2 INTEGER) RETURN NUMBER IS
BEGIN
  RETURN CASE
    WHEN val1 IS NULL OR val2 IS NULL THEN NULL
    WHEN val1 = val2 THEN 1
    ELSE
      0
    END;
END RULE_gebland;
```

Na optimalisatie van de kennisregel waren er meer dan drie uitkomsten mogelijk en moest de gelijkheid uitgelezen worden uit de reconciliatietabel van landen:

[10] Reconciliatiestatistieken van andere knopen dan de centrumknoop gaan dan op dit moment nog verloren. Dit is nog een verbeterpunt voor het algoritme.

```
FUNCTION RULE_geblandn(LAND1 INTEGER, LAND2 INTEGER) RETURN NUMBER
IS
  sim NUMBER;
  TYPE ProbCurTyp IS REF CURSOR;
  prob_cv ProbCurTyp;
BEGIN
  IF (LAND1 IS NULL) OR (LAND2 IS NULL) THEN
    sim := NULL;
  ELSIF (LAND1=LAND2) THEN
    sim := 1;
  ELSE

    OPEN prob_cv FOR
      'SELECT SIM ' ||
      'FROM LANDSIM ' ||
      'WHERE LANDID1=:pk1 AND LANDID2=:pk2'
    USING LAND1, LAND2;

    FETCH prob_cv INTO sim;
    IF prob_cv%NOTFOUND THEN
      sim := 0;
    END IF;

  END IF;
  CLOSE prob_cv;

  RETURN sim;
END RULE_geblandn;
```

E.3.4 Kennisregel ‘pleegdatum’

```
FUNCTION RULE_antdatum(val1 DATE, val2 DATE) RETURN NUMBER IS
  diff INTEGER;
  sim NUMBER;
BEGIN
  IF (val1 IS NULL) OR (val2 IS NULL) THEN
    sim := NULL;
  ELSE
    -- antdatum(HKS) = datum pv >= pleegdatum(OMDATA)
    sim := 0;
    diff := val1-val2;
    IF (diff >= 0) AND (diff <= 20) THEN
      sim := exp(-diff*diff/50);
    END IF;
    IF (diff > 366) THEN
      -- Cutting off pv's with a difference of more than a year can
      -- cause misses. In this case performance is choosen above
      -- correctness at pv level; it is expected to have a negligible
      -- effect on the correctness of the matching at persoon level.
      sim := 0;
    ELSE
      IF (sim <= [EROS::THRESHOLD]) THEN
        sim := [EROS::THRESHOLD]+0.01;
      END IF;
    END IF;
  END IF;

  RETURN sim;
END RULE_antdatum;
```

E.3.5 Kennisregel ‘wetsartikelen’

Het reconciliëren van de wetsartikelen is gebeurd vóór de persoonsreconciliatie. Zie voor een uitgebreide beschrijving van dit proces de bijlage ‘Reconciliatie wetsartikelen’. De gelijkheid wordt uitgelezen uit de reconciliatietabel van wetsartikelen:

```
FUNCTION RULE_pvwa(WA1 INTEGER, WA2 INTEGER) RETURN NUMBER
IS
  sim NUMBER;
  TYPE ProbCurTyp IS REF CURSOR;
  prob_cv ProbCurTyp;
BEGIN
```

```
  OPEN prob_cv FOR
    'SELECT SIMILARITY ' ||
    'FROM EROSS_WA ' ||
    'WHERE PK_HKS=:pk1 AND PK_OMDATA=:pk2'
  USING WA1, WA2;

  FETCH prob_cv INTO sim;
  IF prob_cv%NOTFOUND THEN
    sim := 0;
  END IF;
  CLOSE prob_cv;

  RETURN sim;
END RULE_pvwa;
```

Deze kennisregel kan geen waarde ‘niet beschikbaar’ opleveren, omdat wetsartikel altijd ingevuld is. Voor deze kennisregel is een geoptimaliseerde samenvoegfunctie geschreven:

```
FUNCTION MERGE_wa(simarray SimilarityArray, arraytype INTEGER, sumocc1 INTEGER, sumocc2
INTEGER) RETURN NUMBER IS
  sim NUMBER;
  sim1 NUMBER;
  sim2 NUMBER;
BEGIN
  sim := NULL;
  IF (simarray.COUNT > 0) THEN
     StoreSimilarities(simarray, sumocc1, sumocc2);

     -- total maximum similarity in source 1
     EXECUTE IMMEDIATE
        'SELECT SUM(MAXSIM) FROM (' ||
        '  SELECT MAX(SIMILARITY) AS MAXSIM ' ||
        '  FROM EROST_PROB ' ||
        '  GROUP BY PK_HKS ' ||
        ') source1'
     INTO sim1;
     IF (sim1 IS NULL) THEN sim1 := 0; ELSE sim1 := sim1 / sumocc1; END IF;

     -- total maximum similarity in source 2
     EXECUTE IMMEDIATE
        'SELECT SUM(MAXSIM) FROM (' ||
        '  SELECT MAX(SIMILARITY) AS MAXSIM ' ||
        '  FROM EROST_PROB ' ||
        '  GROUP BY PK_OMDATA ' ||
        ') source1'
     INTO sim2;
     IF (sim2 IS NULL) THEN sim2 := 0; ELSE sim2 := sim2 / sumocc2; END IF;

     -- select minimum similarity
     IF (sim1 < sim2) THEN sim := sim1; ELSE sim := sim2; END IF;

  ELSE
     EXECUTE IMMEDIATE 'DELETE FROM EROST_PROB';
  END IF;

  RETURN sim;
END MERGE_wa;
```

Bijlage F Presentatie resultaten

De resultaten zijn ingedeeld zoals besproken in hoofdstuk 6:

		Gelijk in werkelijkheid?	
		ja	**nee**
Gelijk in EROS?	**ja**	goed-positief	fout-positief
	nee	fout-negatief	goed-negatief

Tabel Bijlage F.1 – Contingentietabel voor resultaten

reden fout	**fout-positief**	**fout-negatief**
niet aanwezig	onterecht gekozen	-[11]
niet gekozen	verkeerd gekozen	niet gekozen
niet gevonden	niet gevonden (fout-positief)	niet gevonden (fout-negatief)

Tabel Bijlage F.2 – Indeling incorrecte resultaten

Voor persoonsentiteiten in OMDATA is er altijd een reconciliatiemogelijkheid; de categorieën 'goed-negatief' en 'onterecht gekozen' zijn voor OMDATA dan ook altijd leeg. Hierdoor is het aantal persoonsentiteiten in de goede resultaten bovendien voor beide bronnen gelijk.

De resultaten worden op twee manieren gepresenteerd: in cijfers en in grafieken. Bij de presentatie in cijfers staat kort genoemd welke aanpassingen er voor de ronde gedaan zijn.

[11] Dit is een correct resultaat, namelijk correct-negatief.

F.1 Presentatie in cijfers

Ronde 1

	HKS	OMDATA
Goed	87,8%	87,1%
Fout	12,2%	12,9%

Tabel Bijlage F.3 – Ronde 1 – Behaalde resultaten

		HKS				OMDATA			
		Gelijk in werkelijkheid?				*Gelijk in werkelijkheid?*			
		ja		**nee**		**ja**		**nee**	
Gelijk in EROS?	**ja**	7585	75,9%	380	3,8%	7585	87,1%	284	3,3%
	nee	836	8,4%	1199	12,0%	836	9,6%	0	0,0%

Tabel Bijlage F.4 – Ronde 1 – Contingentietabel van de behaalde resultaten

Categorie	**HKS**		**OMDATA**	
Niet gevonden (fout-negatief)	270	22,2%	253	22,6%
Niet gevonden (fout-positief)	16	1,3%	33	2,9%
Niet gekozen	110	9,0%	31	2,8%
Verkeerd gekozen	724	59,5%	803	71,7%
Onterecht gekozen	96	7,9%	0	0,0%

Tabel Bijlage F.5 – Ronde 1 – Uitsplitsing foute resultaten

In de eerste ronde zijn de volgende kennisregels geïmplementeerd (details in Bijlage E):

- geboortedatum
- geslacht
- geboorteland
- antecedentdatum versus pleegdatum
- wetsartikelen (zie ook D.2)

Uit de cijfers is af te lezen dat het aantal verkeerd of onterecht gekozen mogelijkheden de grootste groep vormt in de foute resultaten. Dit duidt erop dat de werkelijkheid nog niet goed beschreven wordt of de gemeenschappelijke eigenschappen niet identificerend genoeg zijn. Na analyse van de resultaten blijkt dat de gelijkheid tussen wetsartikelen niet goed beschreven wordt. Dit is in de tweede ronde verbeterd.

Ronde 2

	HKS	OMDATA
Goed	93,2%	87,1%
Fout	6,8%	12,9%

Tabel Bijlage F.6 – Ronde 1 – Behaalde resultaten

		HKS				OMDATA			
		Gelijk in werkelijkheid?				*Gelijk in werkelijkheid?*			
		ja		**nee**		**ja**		**nee**	
Gelijk in EROS?	**ja**	8048	80,5%	538	5,4%	8048	92,5%	547	6,3%
	nee	142	1,4%	1272	12,7%	110	1,3%	0	0,0%

Tabel Bijlage F.7 – Ronde 2 – Contingentietabel van de behaalde resultaten

Categorie	**HKS**		**OMDATA**	
Niet gevonden (fout-negatief)	508	74,7%	524	79,8%
Niet gevonden (fout-positief)	28	4,1%	12	1,8%
Niet gekozen	30	4,4%	23	3,5%
Verkeerd gekozen	91	13,4%	98	14,9%
Onterecht gekozen	23	3,4%	0	0,0%

Tabel Bijlage F.8 – Ronde 2 – Uitsplitsing foute resultaten

In de tweede ronde zijn de volgende verbeteringen doorgevoerd (details in Bijlage E):

- speciale reconciliatiemethode voor wetsartikelen
- introductie geboortejaar in HKS
- introductie maximumafstand tussen antecedentdatum en pleegdatum (1 jaar)

De verbeteringen in deze ronde hebben geleid tot de gewenste daling in verkeerd of onterecht gekozen mogelijkheden. Een deel van de correcte mogelijkheden wordt nu echter ook niet gevonden. Deze categorie vormt nu de grootste groep in de foute resultaten. In de derde ronde wordt geprobeerd deze mogelijkheden te vinden. De belangrijkste oorzaak blijkt echter te liggen in inconsistentie attribuutwaarden en missende delictgegevens (zie hoofdstuk 6). Alleen door de vergelijking van landen te verruimen, kunnen nog betere resultaten worden behaald. De vergelijking van landen is verantwoordelijk voor slechts 117 niet gevonden mogelijkheden. Alleen deze mogelijkheden maken nog een kans door een verbetering gevonden te worden.

Ronde 3

	HKS	OMDATA
Goed	93,8%	93,1%
Fout	6,2%	6,9%

Tabel Bijlage F.9 – Ronde 3 – Behaalde resultaten

		HKS				OMDATA			
		Gelijk in werkelijkheid?				*Gelijk in werkelijkheid?*			
		ja		**nee**		**ja**		**nee**	
Gelijk in EROS?	**ja**	8108	81,1%	133	1,3%	8108	93,1%	133	1,5%
	nee	490	4,9%	1269	12,7%	464	5,3%	0	0,0%

Tabel Bijlage F.10 – Ronde 1 – Contingentietabel van de behaalde resultaten

Categorie	**HKS**		**OMDATA**	
Niet gevonden (fout-negatief)	459	4,6%	440	5,1%
Niet gevonden (fout-positief)	11	0,1%	30	0,3%
Niet gekozen	31	0,3%	24	0,3%
Verkeerd gekozen	96	1,0%	103	1,2%
Onterecht gekozen	26	0,3%	0	0,0%

Tabel Bijlage F.11 – Ronde 3 – Uitsplitsing foute resultaten

In de derde ronde is de volgende verbetering doorgevoerd (details in Bijlage D):

- nieuwe reconciliatie van landen met partiële gelijkheid (Nederland en Nederlandse Antillen zijn nu bijvoorbeeld 0.8 gelijk in plaats van 0)

Doordat in deze ronde slechts één verbetering is doorgevoerd, kan precies worden nagegaan wat de effecten van deze verbetering zijn geweest door deze resultaten te vergelijken met de resultaten uit de vorige ronde.

	HKS			**OMDATA**		
Categorie	*2e ronde*	*3e ronde*	*verschil*	*2e ronde*	*3e ronde*	*verschil*
Goed-positief	8048	8108	60	8048	8108	60
Goed-negatief	1272	1269	-3	0	0	0
Niet gevonden (fout-negatief)	508	459	-49	524	440	-84
Niet gevonden (fout-positief)	28	11	-17	12	30	18
Niet gekozen	30	31	1	23	24	1
Verkeerd gekozen	91	96	5	98	103	5
Onterecht gekozen	23	26	3	0	0	0

Tabel Bijlage F.12 – Verschil tussen de 2e en 3e ronde

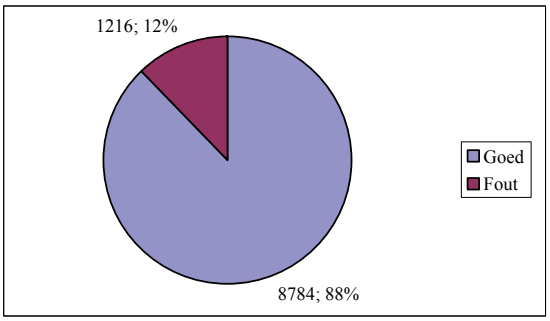
Het aantal mogelijkheden dat niet gevonden wordt is per saldo afgenomen met 66. In de tweede ronde was de vergelijking van geboortelanden nog verantwoordelijk voor 117 niet gevonden mogelijkheden. In die zin is de verbetering geslaagd: voor meer dan de helft is nu een reconciliatiemogelijkheid gevonden (56%). De schaduwzijde is dat door de verbetering ook (nieuwe) verkeerde keuzes worden gemaakt of oude goede keuzes niet meer worden gemaakt. Desondanks is het resultaat per saldo positief (57 in HKS, 60 in OMDATA) en liggen de aantallen dicht bij het aantal nieuw gevonden mogelijkheden. De verbetering is daarmee geslaagd.

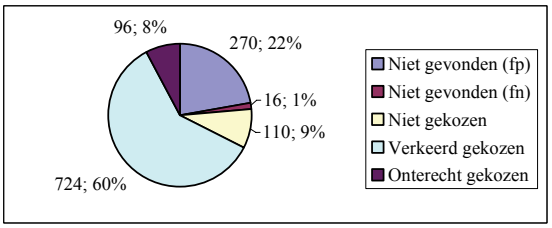
F.2 Presentatie in grafieken

Deze paragraaf presenteert de resultaten in grafiekvorm. Hierdoor worden de verhoudingen tussen de verschillende waarden sneller zichtbaar. De grafieken bevatten geen nieuwe informatie.

Ronde 1

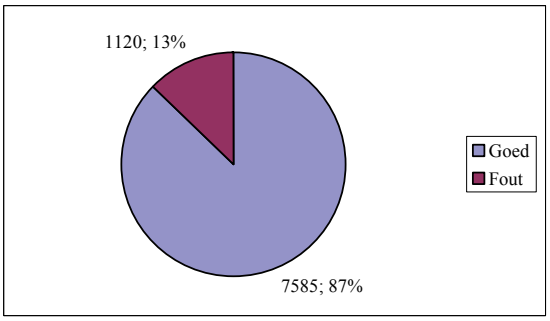
Vanuit HKS

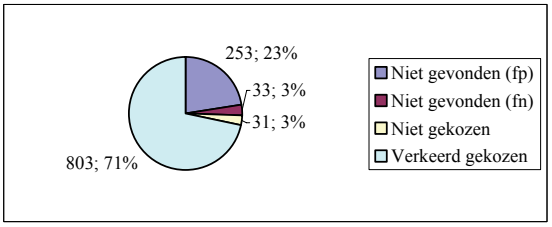




Figuur Bijlage F.1 (l) – Ronde 1 – HKS – Behaalde resultaten [n=10.000]
Figuur Bijlage F.2 – Ronde 1 – HKS – Uitsplitsing foute resultaten [n=1.216]

Vanuit OMDATA

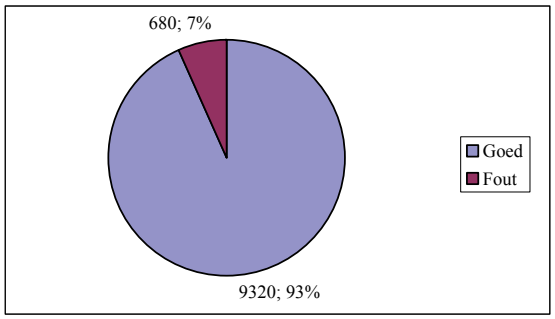


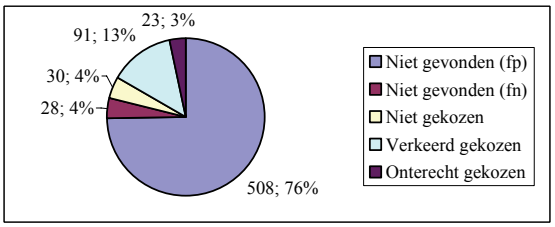


Figuur Bijlage F.3 (l) – Ronde 1 – OMDATA – Behaalde resultaten [n=8.705]
Figuur Bijlage F.4 (r) – Ronde 1 – OMDATA – Uitsplitsing foute resultaten [n=1.120]

Ronde 2

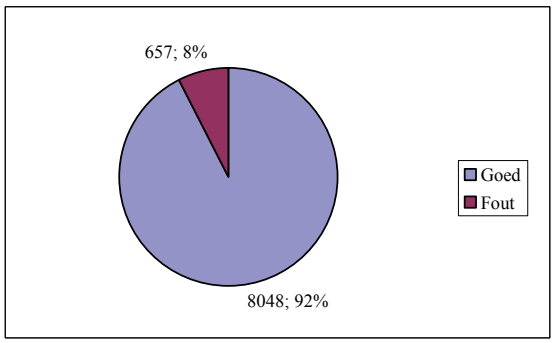
Vanuit HKS

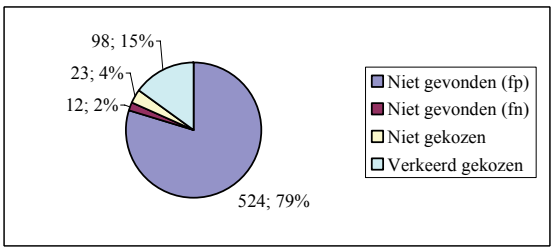




Figuur Bijlage F.5 (l) – Ronde 2 – HKS – Behaalde resultaten [n=10.000]
Figuur Bijlage F.6 (r) – Ronde 2 – HKS – Uitsplitsing foute resultaten [n=680]

Vanuit OMDATA

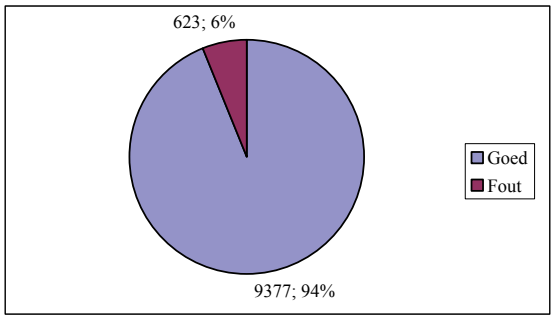


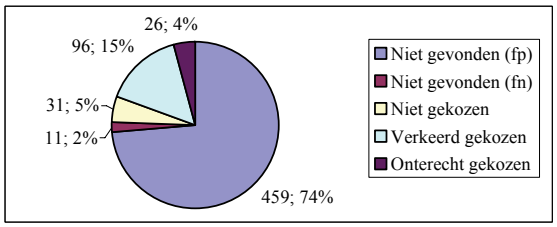


Figuur Bijlage F.7 (l)– Ronde 2 – OMDATA – Behaalde resultaten [n=8.705]
Figuur Bijlage F.8 (r) – Ronde 2 – OMDATA – Uitsplitsing foute resultaten [n=657]

Ronde 3

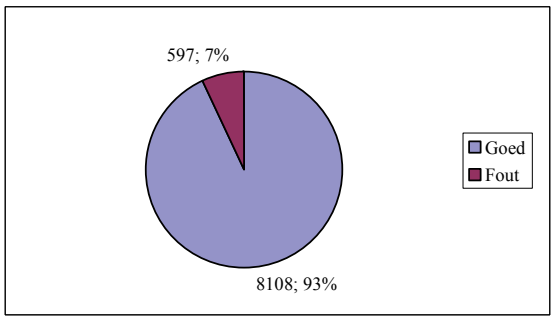
Vanuit HKS

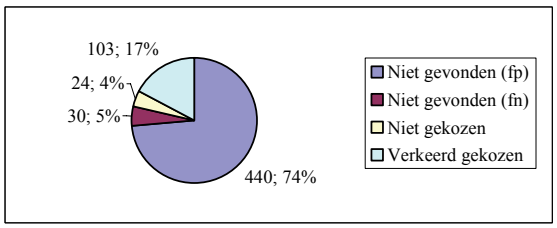




Figuur Bijlage F.9 (l) – Ronde 3 – HKS – Behaalde resultaten [n=10.000]
Figuur Bijlage F.10 (r) – Ronde 3 – HKS – Uitsplitsing foute resultaten [n=623]

Vanuit OMDATA





Figuur Bijlage F.11 (l) – Ronde 3 – OMDATA – Behaalde resultaten [n=8.705]
Figuur Bijlage F.12 (r) – Ronde 3 – OMDATA – Uitsplitsing foute resultaten [n=597]

Bijlage G Extrapolatie

In dit onderzoek is een representatieve steekproef uit twee informatiebronnen gebruikt als referentieset. Van deze referentieset waren de correcte reconciliaties bekend, waardoor het mogelijk was de correctheid en de kwaliteit van de gemaakte reconciliaties te bepalen. Is dit ook mogelijk voor een andere (grotere) steekproef, of misschien zelfs voor de volledige informatiebronnen, zonder dat de correcte reconciliaties bekend zijn?

De persoonsreconciliatie is uitgevoerd met behulp van kennisregels. Deze kennisregels beschrijven hoe eigenschappen van twee entiteiten uit verschillende bronnen zich ten opzichte van elkaar gedragen als ze tot hetzelfde object zouden behoren. De kennisregels gelden dus ook als er andere steekproef wordt genomen, of zelfs de volledige informatiebronnen, zolang de kennisregels maar dezelfde objecten blijven beschrijven.

Als de kennisregels een persoonsentiteit volledig kunnen identificeren, dan is er geen probleem. Als de kennisregels echter niet genoeg discrimineren, dan ontstaan er meerdere reconciliatiemogelijkheden per persoonsentiteit. Hoe meer reconciliatiemogelijkheden, hoe groter de kans dat de verkeerde reconciliatiemogelijkheid wordt gekozen.

Door de referentieset zodanig te kiezen dat het aantal reconciliatiemogelijkheden in de referentieset vergelijkbaar in aantal is met het aantal reconciliatiemogelijkheden in de onbekende set, kan het gedrag van het reconciliatieproces voorspelt worden.

Het precieze aantal reconciliatiemogelijkheden per entiteit is pas bekend nadat het reconciliatieproces is uitgevoerd. Vanwege de lange duur van dit proces is dit niet wenselijk. Het aantal mogelijkheden kan ook afgeschat worden.

We selecteren de persoonsattributen die, afgezien van inconsistenties en missende informatie, een sterke vergelijking (gelijk of niet) vormen. Dit zijn geboortedatum, geslacht en geboorteland. Deze attributen vormen samen de *extrapolatiesleutel*.

Indien de extrapolatiesleutel uniek is in beide informatiebronnen, dan is er geen probleem. De entiteiten worden gereconcilieerd als er voldoende gelijkheid is. Hoe meer entiteiten dezelfde extrapolatiesleutel hebben, hoe kleiner de kans dat de juiste reconciliatiemogelijkheid gekozen wordt.

De aantallen bij elke extrapolatiesleutel zijn te berekenen voor zowel de referentieset als de onbekende set. Door ervoor te zorgen dat de aantallen in de referentieset overeenkomen met de aantallen in de onbekende set, kunnen de kansen worden geëxtrapoleerd.

Bij een groot verschil in aantal entiteiten tussen de referentieset en de onbekende set, is het goed mogelijk dat de extrapolatiesleutel-statistieken niet overeenkomen. Door de extrapolatiesleutel te vervagen, kan dit doel toch bereikt worden. Zo kan bijvoorbeeld de geboortemaand in de sleutel worden opgenomen in plaats van de geboortedatum. Als de statistieken eenmaal redelijk overeenkomen, dan kan de referentieset met de vervaging gereconcilieerd worden. De behaalde resultaten zijn nu extrapoleerbaar naar de onbekende set.

Alleen voor attributen op persoonsniveau of erboven heeft het zin om ze te vervagen. Attributen onder persoonsniveau gelden immers altijd per persoon, en die verhouding blijft gelijk.

Box Bijlage G.1 – Extrapolatie in de onderzoeksomgeving. In dit onderzoek is gebruik gemaakt van een referentieset met gegevens uit twee informatiebronnen HKS en OMDATA. Voor het WODC is het erg interessant hoe de huidige set van kennisregels extrapoleert naar deze informatiebronnen als geheel.

OMDATA heeft op dit moment het probleem dat er geen persoonsniveau beschikbaar is. De extrapolatiesleutels kunnen wel berekend worden, maar het is niet mogelijk om het aantal reconciliatiemogelijkheden af te schatten.

Verder onderzoek binnen het WODC kan wellicht aantonen of via andere wegen het extrapolatiegedrag toch bepaald kan worden. Bijvoorbeeld door alleen het aantal reconciliatiemogelijkheden uit HKS te gebruiken, waarbij aangetoond moet worden dat HKS in die zin representatief is voor OMDATA. Of door het extrapolatiegedrag te voorspellen op proces-verbaalniveau en ook de referentieset op dit niveau te conciliëren.

Box Bijlage G.1 – Extrapolatie in de onderzoeksomgeving

Bijlage H Formele theorie

In deze bijlage is de theorie uit hoofdstuk 3 formeel uitgewerkt. Indien van toepassing is verwezen naar verwante definities en andere passages in de hoofdtekst.

Definitie 1. Een object is iets uit de reële wereld.

Definitie 2. Een entiteit is een (deel)representatie van een object; "object *o* wordt gerepresenteerd door entiteit *e*" wordt genoteerd als "$[[e]] = o$".

Definitie 3. Laat A een stel functies zijn (notatie: $a(x) \equiv x.a$). Dan is O een A-objectsoort als:

O is een verzameling objecten zdd (zodanig dat) $\forall a \in A, o \in O : o \in \operatorname{dom} a$ ($a(o)$ is zinvol)

Definitie 4. Laat A een stel functies zijn, en O een A-objectsoort. Dan is E een entiteitsoort voor O als:

E is een verzameling entiteiten zdd $\forall e \in E : [[e]] \in O$

Definitie 5. Laat A een stel functies zijn, en O een A-objectsoort. Laat verder E een entiteitsoort voor O zijn. Dan is $a \in A$ een attribuut van E als:

$\forall e \in E : e \in \operatorname{dom} a$ ($a(e)$ is zinvol)

Opmerking. Zie Intuïtie 1 voor de mogelijkheid dat $e.a \neq [[e]].a$.

Verwijzing. Buiten deze bijlage wordt gesproken over 'entiteittype' in plaats van de bredere definitie 'entiteitsoort'.

Definitie 6. Laat E een entiteitsoort voor (objectsoort) O, en s een attribuut van E (en dus O) zijn. Dan is s een sleutelattribuut als:

1. $\forall e \in E : e.s = [[e]].s$
2. $\forall e, e' \in E : e.s = e'.s \Rightarrow e = e'$

Opmerking. Een alternatief voor de tweede eigenschap zou kunnen zijn:
$\forall o, o' \in O : o.s = o'.s \Rightarrow o = o'$. Dit alternatief wordt verworpen vanwege:

- "Attribuut" slaat op entiteiten en niet op objecten.
- Sommige $a \in A$ worden artificieel geïntroduceerd aan de hand van entiteiten.

Eis 1. Elke entiteitsoort heeft een sleutelattribuut.

Verwijzing. In hoofdstuk 3 wordt de eis gesteld dat elke entiteit identificeerbaar is aan de hand van een sterke sleutel (pagina 10, onderaan). Een sterke sleutel kan bestaan uit meerdere attributen, in plaats van één enkel sleutelattribuut. In dit geval wordt het sleutelattribuut artificieel geïntroduceerd aan de hand van de sterke sleutel.

Postulaat 1. Gegeven een *A*-objectsoort. Laat $\vec{a} \subseteq A$ en $\vec{b} \subseteq A$. De functie

$$positie_{\vec{a},\vec{b}}(\vec{a},\vec{b})$$

positioneert de waarde(n) van $\vec{b}$ ten opzichte van waarde(n) van $\vec{a}$, op een lijn van -∞..∞ met de waarde(n) van $\vec{a}$ op positie 0, zdd een positie dichter bij 0 een grotere mate van overeenkomst geeft. Als positionering niet mogelijk is, dan geeft de functie als uitkomst "n.a." (*not available*).

Opmerking. De functie *positie* wordt genoteerd met $\vec{a}, \vec{b}$ als subscript en argument, waarmee respectievelijk de namen en waarden van $\vec{a}$ en $\vec{b}$ bedoeld worden. Deze notatie is informeel, maar goed genoeg voor ons doel.

Verwijzing. Buiten deze bijlage wordt gesproken over 'afstand' in plaats van 'positie'. Dit is niet correct, omdat 'positie' de volgende extra eigenschappen heeft;

- mogelijk negatief: $positie_{\vec{a},\vec{b}}(\vec{a},\vec{b}) < 0$;
- mogelijk asymmetrisch: $\left|positie_{\vec{a},\vec{b}}(\vec{a},\vec{b})\right| \neq \left|positie_{\vec{b},\vec{a}}(\vec{b},\vec{a})\right|$.

Intuïtie 1. De relatie tussen attributen van *e*, *o*. Als [[*e*]] = *o*, dan:

1. Voor veel (maar niet noodzakelijk alle) attributen *a* geldt: *e.a* = *o.a*.
2. Als geldt *e.a* ≠ *o.a*, dan verwachten we "*positie*(*e.a*, *o.a*) is klein".

Definitie 7. De mate van overeenkomst drukken we uit in een *similarity*-waarde:

$$[0..1] \cup \{\text{"}n.a.\text{"}\}$$

Verwijzing. Buiten deze bijlage wordt de *similarity*-waarde een gelijkheidswaarde genoemd (zie definitie op pagina 11).

Definitie 8. "Gerepresenteerd door", [[*x*]] = *y*

"*y* uit de reële wereld wordt gerepresenteerd door *x*", in andere woorden "de interpretatie van *x* is *y*" als volgt gedefinieerd:

1. entiteit, object: $[[e]] = o \equiv e.s = o.s$
2. entiteitsoort, objectsoort: $[[E]] = \{e : E \bullet [[e]]\}$

Definitie 9. Laat E_1, E_2 entiteitsoorten voor *A*-objectsoort *O* zijn. Een gemeenschappelijke eigenschap voor E_1, E_2 is een drietal ($\vec{a}_1$, $\vec{a}_2$, *freq*), zdd:

1. $e_1.\vec{a}_1$ bestaat voor $\forall e_1 \in E_1$
2. $e_2.\vec{a}_2$ bestaat voor $\forall e_2 \in E_2$
3. *freq* is een functie zdd:

$$freq(d) = \frac{\#\{o \in O \mid positie_{\vec{a}_1,\vec{a}_2}(\vec{a}_1,\vec{a}_2) = d\}}{\#O}, \text{met } d \in \text{ran } positie_{\vec{a}_1,\vec{a}_2} - \{"n.a."\}$$

Opmerking. De functie *freq* heeft een bultvorm door de eigenschappen van de functie *positie* (zie Postulaat 1).

Verwijzing. De functie *freq* wordt in hoofdstuk 3 frequentiefunctie genoemd (zie pagina 12, 2e definitie).

Definitie 10. Laat $p = (\vec{a}_1, \vec{a}_2, freq)$ een gemeenschappelijke eigenschap voor E_1, E_2 zijn. De *similarity* in de gemeenschappelijke eigenschap p

$sim_p : E_1 \times E_2 \rightarrow$ *similarity*-waarde

wordt gedefinieerd als:

$$sim_p(e_1,e_2) = \begin{cases} \dfrac{freq(positie_{\vec{a}_1,\vec{a}_2}(e_1.\vec{a}_1, e_2.\vec{a}_2))}{\max(\text{ran } freq)} & \text{als } positie_{\vec{a}_1,\vec{a}_2}(e_1.\vec{a}_1, e_2.\vec{a}_2) \neq "n.a." \\ "n.a." & \text{anders} \end{cases}$$

Verwijzing. In hoofdstuk 3 wordt de *similarity* in een gemeenschappelijke eigenschap een gelijkheidsfunctie genoemd (3e definitie op pagina 12).

Definitie 11. Laat A een verzameling functies zijn, en E_1 en E_2 entiteitsoorten voor een zelfde A-objectsoort. Laat $p_i = (\vec{a}_{i1}, \vec{a}_{i2}, freq_i)$ $(i = 1..n)$ gemeenschappelijke eigenschappen voor E_1, E_2 zijn, met verzamelingen D_{i1} en D_{i2} zdd:

$$\forall e_1 : E_1, e_2 : E_2 \forall i, j \in \{1,..,n\}, i \neq j \bullet \neg \quad ((e_1.\vec{a}_{i1} \in D_{i1} \wedge e_2.\vec{a}_{i2} \in D_{i2}) \wedge (e_1.\vec{a}_{j1} \in D_{j1} \wedge e_2.\vec{a}_{j2} \in D_{j2}))$$

($\langle D_{i1} \times D_{i2} \rangle$ is een (deel)partitionering van ran$(\vec{a}_{i1}, \vec{a}_{i2})$)

De *similarity* in de disjuncte gemeenschappelijke eigenschappen $p_1..p_n$

$sim_{(p_1,..,p_n)} : E_1 \times E_2 \rightarrow$ *similarity*-waarde

wordt gedefinieerd als:

$$sim_{(p_1,..,p_n)}(e_1,e_2) = \begin{cases} sim_{p_1}(e_1,e_2) & \text{als } e_1.\vec{a}_{11} \in D_{11} \wedge e_2.\vec{a}_{12} \in D_{12} \\ \vdots & \vdots \\ sim_{p_n}(e_1,e_2) & \text{als } e_1.\vec{a}_{n1} \in D_{n1} \wedge e_2.\vec{a}_{n2} \in D_{n2} \\ "n.a." & \text{anders} \end{cases}$$

We noemen $sim_{(p_1,..,p_n)}$ een *A*-kennisregel voor E_1, E_2.

Verwijzing. In hoofdstuk 3 (definitie op pagina 14) wordt $sim_{(p_1,..,p_n)}$ gedefinieerd met $\vec{a}_{i1} = \vec{a}_{j1}$ en $\vec{a}_{i2} = \vec{a}_{j2}$ (voor alle *i*, *j*). Verder wordt de functie genoteerd met behulp van een omschrijving; $sim_{omschrijving}$.

Definitie 12. Laat $x_1..x_n$ *similarity*-waarden zijn (dus in $[0..1] \cup \{"n.a."\}$). De *n*-aire *similarity*-operator wordt gedefinieerd als:

$$x_1 \oplus ... \oplus x_n = \begin{cases} 0 & \text{als } \exists i : x_i = 0 \\ "n.a." & \text{als } \forall i : x_i = "n.a." \\ \dfrac{\sum_{i:x_i \neq "n.a."} x_i}{\#\{i \mid x_i \neq "n.a."\}} & \text{anders} \end{cases}, \text{ met } i \in \{1,..,n\}$$

Opmerking. Het symbool $\oplus$ is suggestief. De *similarity*-operator heeft echter zowel eigenschappen van een optelling, als van een vermenigvuldiging. Het symbool $\otimes$ is daarom ook een alternatief.

Verwijzing. In hoofdstuk 3 wordt de *similarity*-operator gelijkheidsoperator genoemd en wordt het symbool $\circ$ gebruikt (definitie gelijkheidsoperator op pagina 22). De stelling op pagina 23 toont aan dat de *similarity*-operator commutatief en "niet-associatief" is.

Definitie 13. Laat E_1, E_2 entiteitsoorten voor *A*-objectsoort *O* zijn. Laat $sim_1..sim_m$ *A*-kennisregels voor E_1, E_2 zijn ($sim_{(p_{i1},..,p_{in})}$ wordt afgekort tot sim_i). De *similarity* tussen entiteiten

$$s_O : E_1 \times E_2 \rightarrow \textit{similarity}\text{-waarde}$$

wordt gedefinieerd als:

$$s_O(e_1,e_2) = (sim_1 \oplus ... \oplus sim_m)(e_1,e_2)$$

Opmerking. Pas op: s_O hangt af van het gekozen stel kennisregels. Wanneer we over s_O spreken, dan wordt verondersteld dat het stel kennisregels bekend is (en vaststaat). Vollediger zou zijn om het stel kennisregels als parameter mee te geven, bijvoorbeeld als subscript aan "*s*".

Verwijzing. Deze definitie is gelijk aan de definitie van entiteitgelijkheid op pagina 23 zonder gewichten voor kennisregels.

Definitie 14. Een ER-diagram is een tweetal (*Es*, *Rs*) zdd:

1. *Es* is een verzameling van entiteitsoorten.
2. *Rs* is een verzameling van relaties tussen entiteitsoorten.

Eis 2. Relaties zijn binair; $R \in Rs, E_1 \in Es, E_2 \in Es : R \subseteq E_1 \times E_2$.

Verwijzing. Relaties met een hogere graad en andere gevallen uit de E/R modellering die niet in bovenstaande definitie passen, worden nader toegelicht in hoofdstuk 3 (pagina 19, onderaan).

Definitie 15. Laat D_1, D_2 ER-diagrammen zijn, $D_1 = (Es_1, Rs_1)$ en $D_2 = (Es_2, Rs_2)$. Het verenigd ER-diagram *V* van D_1 en D_2 is een tweetal (*Es*, *Rs*) zdd:

1. $Es = \{(E_1, E_2) \mid E_1 \in Es_1 \wedge E_2 \in Es_2 \wedge$
   ("[[E_1]] en [[E_2]] zijn beide *A*-objectsoorten voor een overeenkomstige *A*")$\}$

2. $Rs = \{(R_1, R_2) \mid (R_1 \in Rs_1 \wedge R_1 \subseteq E_{11} \times E_{21}) \wedge (R_2 \in Rs_2 \wedge R_2 \subseteq E_{12} \times E_{22}) \wedge$
   $(E_{11}, E_{21}) \in Es_1 \wedge (E_{12}, E_{22}) \in Es_2\}$

Definitie 16 (wiskundig). Een rooted-DAG is een drietal (*Kn*, *Pn*, *wortel*) zdd:

1. *Kn* is een verzameling van knopen.
2. *Pn* is een verzameling van pijlen.

   Elke pijl *P* heeft een beginpunt, genoteerd als $bron(P) \in Kn$, en een eindpunt, genoteerd als $doel(P) \in Kn$. We noteren $P \in Pn$ met *bron*(*P*) = K_1 en *doel*(*P*) = K_2 als $K_1 \xrightarrow{P} K_2$.
   Verder: $K_1 \rightarrow K_2 \equiv (\exists P : K_1 \xrightarrow{P} K_2)$.

3. $wortel \in Kn$, zdd: $\forall k \in Kn : k \rightarrow^* wortel$
4. Er zijn geen cykels: $\forall k \in Kn : \neg(k \rightarrow^+ k)$

Definitie 17. Laat *V* een ER-diagram zijn, *V* = (*Es*, *Rs*). Laat *G* een rooted-DAG zijn, *G* = (*Kn*, *Pn*, *wortel*). We noemen *G* een gelijkheidsdistributie op *V* als $Kn \subseteq Es \wedge Pn \subseteq Rs$.

Verwijzing. Deze definitie komt overeen met de definitie van gelijkheidsdistributie op pagina 17.

Definitie 18. Laat D_1, D_2 ER-diagrammen zijn. Laat *V* een ER-diagram zijn, *V* = (*Es*, *Rs*), zdd *V* $\subseteq$ het verenigd ER-diagram van D_1 en D_2. Laat *G* een gelijkheidsdistributie op *V* zijn, *G* = (*Kn*, *Pn*, *wortel*) (dus $Kn \subseteq Es \wedge Pn \subseteq Rs$). Laat *K* = (E_1, E_2)$\in Kn$ een knoop uit *V*, *G*; E_1 en E_2 zijn entiteitsoorten voor een zelfde *A*-objectsoort die we *O* noemen. Laat verder s_O de *similarity* tussen entiteiten van *O* zijn, $s_O = (sim_1 \oplus ... \oplus sim_m)$.

De *similarity* tussen entiteiten en hun gerelateerde entiteiten (*object similarity*)

$S_K : E_1 \times E_2 \rightarrow$ *similarity*-waarde

wordt gedefinieerd als: $\forall (e_1, e_2) \in K$,

$$S_K(e_1, e_2) = sim_1 \oplus \ldots \oplus sim_m \quad \oplus T_P(S_{K' \xrightarrow{P} K}(e_1, e_2)) \quad \vdots \quad \oplus \text{ voor alle } P \in Pn \text{ met } doel(P) = K$$

waarbij:

- $S_{K' \xrightarrow{P} K}$ de *similarity*-bijdrage van entiteiten uit *K*' aan de *object similarity* van entiteiten uit *K* is (zie Definitie 19);
- T_P gepostuleerd wordt voor alle $P \in Pn$.

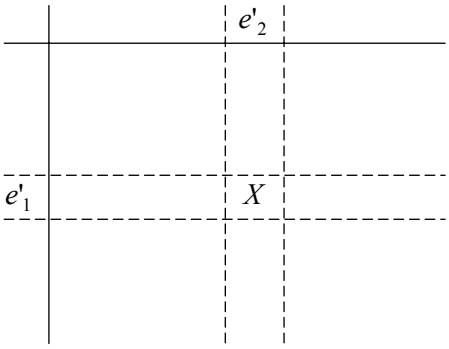
Verwijzing. De definitie voor *object similarity* is gelijk aan de definitie van objectgelijkheid op pagina 24 zonder gewichten voor kennisregels. De transformatiefunctie T_p wordt besproken in paragraaf 3.4.2.

Definitie 19. Laat $K, K' \in Kn$ knopen uit *V*, *G* zijn, *K* = (E_1, E_2). Laat $P \in Pn$ een relatie tussen *K*' en *K* zijn, *P* = (R_1, R_2). De *similarity*-bijdrage van entiteiten uit *K*' aan de *object similarity* van entiteiten uit K

$S_{K' \xrightarrow{P} K} : E_1 \times E_2 \rightarrow$ matrix van *similarity*-waarden

wordt gedefinieerd als: $\forall (e_1, e_2) \in K$,

$S_{K' \xrightarrow{P} K}(e_1, e_2) =$



met $(e'_1, e'_2) \in K'$, $X = S_{K'}(e'_1, e'_2)$, $(e'_1, e_1) \in R_1$ en $(e'_2, e_2) \in R_2$.

Opmerking. Ogenschijnlijk is hier sprake van wederzijdse afhankelijkheid tussen S_K en $S_{K' \xrightarrow{P} K}$, maar vanwege de acycliciteit van *G* is er een *acyclische* gerichte afhankelijkheid.

Definitie 20. Laat $K = (E_1, E_2) \in Kn$ een knoop uit V, G zijn. De *definitieve object similarity* tussen entiteiten

$$S'_K : E_1 \times E_2 \to [0..1]$$

wordt gedefinieerd als: $\forall (e_1, e_2) \in K$,

$$S'_K(e_1, e_2) = \begin{cases} S_K(e_1, e_2) & \text{als } S_K(e_1, e_2) \neq \text{"}n.a.\text{"} \\ 0 & \text{anders} \end{cases}$$

Definitie 21. Laat $K = (E_1, E_2) \in Kn$ een knoop uit V, G zijn. De *waarschijnlijkheid* dat $[[e_1]] = [[e_2]]$ ten opzichte van andere entiteitparen uit E_1, E_2

$$prob_K : E_1 \times E_2 \to [0..1]$$

wordt gedefinieerd als: $\forall (e_1, e_2) \in K$,

$$prob_K(e_1, e_2) = \frac{S'_K(e_1, e_2)}{\sum_{e \in E_2} S'_K(e_1, e)} \cdot \frac{S'_K(e_1, e_2)}{\sum_{e \in E_1} S'_K(e, e_2)}$$

Verwijzing. Deze definitie is gelijk aan de definitie van waarschijnlijkheid op pagina 26.

Intuïtie 2. De kans dat een paar entiteiten (e_1, e_2) hetzelfde object representeren, neemt toe als:
- de mate van overeenkomst toeneemt (*similarity*)
- de mate van overeenkomst in gerelateerde entiteitparen afneemt (*waarschijnlijkheid*)

(gerelateerd: $(\forall e \in E_1, e \neq e_1 : (e, e_2)) \cup (\forall e \in E_2, e \neq e_2 : (e_1, e))$).

In Z-notatie:

Voor willekeurige $a \in [0,1]$ en $pr \in [0,1]$ geldt:

$$\forall e_1 \in E_1 :$$
$$P(e_2 \in E_2 \mid S'_K(e_1, e_2) >> a \vee (a \approx S'_K(e_1, e_2) \wedge prob_K(e_1, e_2) \geq pr) \bullet [[e_1]] = [[e_2]])$$
$$\geq$$
$$P(e'_2 \in E_2 \mid a >> S'_K(e_1, e'_2) \vee (a \approx S'_K(e_1, e'_2) \wedge pr \geq prob_K(e_1, e'_2)) \bullet [[e_1]] = [[e'_2]])$$
$$\forall e_2 \in E_2 :$$
$$P(e_1 \in E_1 \mid S'_K(e_1, e_2) >> a \vee (a \approx S'_K(e_1, e_2) \wedge prob_K(e_1, e_2) \geq pr) \bullet [[e_1]] = [[e_2]])$$
$$\geq$$
$$P(e'_1 \in E_1 \mid a >> S'_K(e'_1, e_2) \vee (a \approx S'_K(e'_1, e_2) \wedge pr \geq prob_K(e'_1, e_2)) \bullet [[e'_1]] = [[e_2]])$$

Opmerking. Doordat objecten binnen één entiteitsoort slechts door één entiteit gerepresenteerd kunnen worden (zie Definitie 6), geldt in bovenstaande vergelijkingen:

$$\forall e_1 \in E_1 : \quad [[e_1]] = [[e_2]] \Rightarrow ([[e_1]] \neq [[e'_2]] \vee e_2 = e'_2)$$
$$\forall e_2 \in E_2 : \quad [[e_1]] = [[e_2]] \Rightarrow ([[e'_1]] \neq [[e_2]] \vee e_1 = e'_1)$$

De operatoren >> en << betekenen respectievelijk “veel groter dan” en “veel kleiner dan”. Er geldt in de vergelijkingen:

$$(a >> S'_K(e_1,e_2) \vee S'_K(e_1,e_2) >> a) \Leftrightarrow \neg(a \approx S'_K(e_1,e_2))$$

Verwijzing. In deze intuïtie speelt de waarschijnlijkheid alleen een rol als de *similarities* dicht bij elkaar liggen. In hoofdstuk 3 zijn *similarity* en waarschijnlijkheid gecombineerd in één waarde; de reconciliatiescore (zie definitie op pagina 26). Hiervoor geldt dezelfde intuïtie als hierboven beschreven.

Intuïtie 3. De kans dat twee entiteiten die *niet* hetzelfde object representeren een grote mate van overeenkomst hebben, neemt af bij een toename van het aantal (gebruikte) gemeenschappelijke eigenschappen.

Intuïtie 2 en Intuïtie 3 zijn – met behulp van een prototype – getoetst door middel van een casus (zie hoofdstuk 4 en 5). Het resultaat is, aan de hand van de probleemstelling en onderzoeksvragen uit hoofdstuk 2, besproken in hoofdstuk 6 en 7.