Memorandum 2009-2

Entiteitreconciliatie ondanks beperkte overlap door middel van objectgelijkenis

Casus 'Koppelen van persoonsgegevens zonder een gemeenschappelijke identificatie'

J.J. van Dijk

Justitie

Wetenschappelijk Onderzoek- en Documentatiecentrum

Exemplaren van deze publicatie kunnen
schriftelijk worden besteld bij

Bibliotheek WODC
Postbus 20301, 2500 EH Den Haag

Fax: (070) 370 45 07
E-mail: wodc@minjus.nl

Memoranda worden in beperkte mate gratis
verspreid zolang de voorraad strekt

Alle nadere informatie over WODC-publicaties
is te vinden op Justweb en op www.wodc.nl

Inhoud

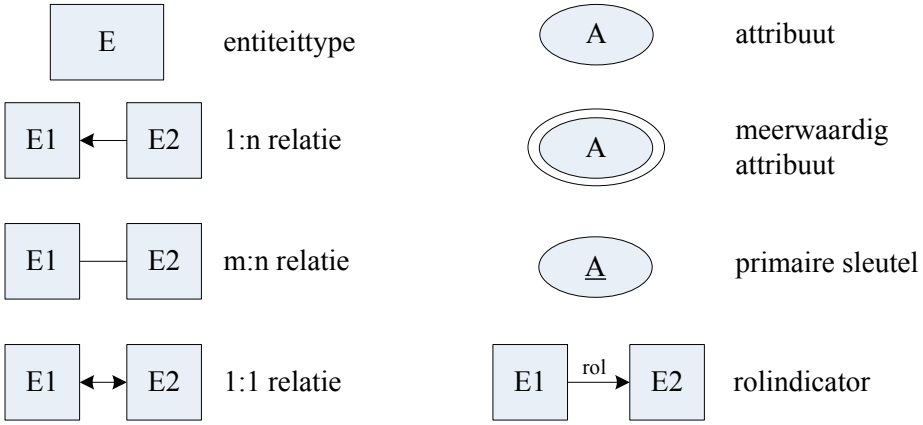
Afkortingen

CBS	Centraal Bureau voor de Statistiek
EROS	Entity Reconciliation using Object Similarity
HKS	HerKenningsdienstSysteem, registratie van processen-verbaal van aangifte van misdrijven
OM	Openbaar Ministerie
OMDATA	Informatiesysteem van het Parket-Generaal van het Openbaar Ministerie
OBJD	Onderzoeks- en Beleidsdatabase Justitiële Documentatie
PV	proces-verbaal

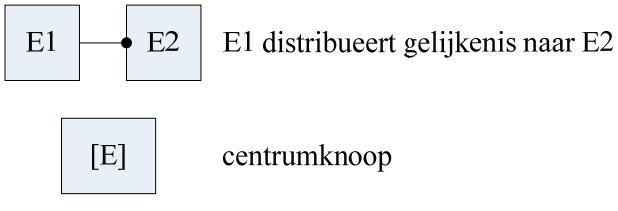
Definities

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 22).
Conciliëren / Conciliatie	Het verenigen van de set van entiteiten in een knoop, waarbij objectgelijke entiteiten gereconcilieerd worden (zie pagina 9).
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 22).
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 9).
Referentieset	Een representatieve subset van één of meer informatiebronnen.

Notaties in datamodellen



Notaties in gelijkenisdistributie



Samenvatting

Het koppelen van informatiebronnen wordt in de huidige maatschappij steeds belangrijker. Door koppelen ontstaan nieuwe inzichten, omdat meer 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 van twee bronnen te gebruiken. Overlap bestaat uit eigenschappen die beide bronnen gemeen hebben. Als een gemeenschappelijke eigenschap overeenkomt, dan is er sprake van gelijkenis (Eng. *similarity*). De mate van gelijkenis wordt bepaald door de onderlinge positionering van twee attributen die de eigenschap beschrijven. Met behulp van expertkennis wordt deze positie via een positieverdeling (een trendlijn over het histogram van de verwachte onderlinge positionering van de eigenschap) omgezet in een gelijkeniswaarde. 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 entiteitgelijkenis per knoop bepaald en wordt ook de objectgelijkenis bepaald, waarin tevens de gelijkenis in andere knopen wordt meegenomen. Hierbij wordt de gelijkenis 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 objectgelijkenis 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 privacygerelateerde toepassingen.

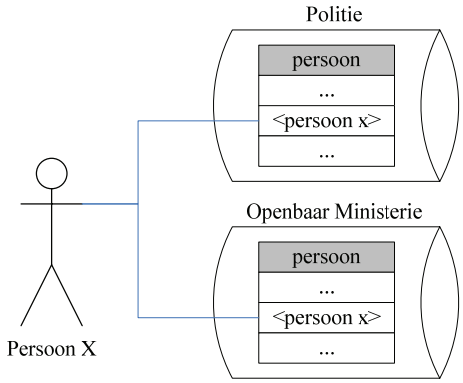
1 Introductie

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).

Figuur 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. (2002) 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 notatie $[[p]]$ het object behorende bij een entiteit p representeert.

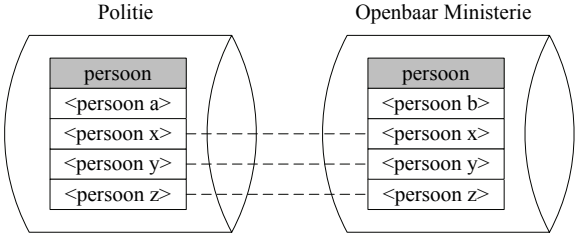
Definitie. De conciliatie van twee sets van entiteiten P_1 en P_2 is de vereniging $P_1 \cup P_2$ waarvoor geldt:

$$\forall p_1 \in P_1 : \quad p_1 \in P_1 \cap P_2 \quad \leftrightarrow \quad \exists! \, p_2 \in P_1 \cap P_2 : [[p_1]] = [[p_2]]$$
$$\forall p_2 \in P_2 : \quad p_2 \in P_1 \cap P_2 \quad \leftrightarrow \quad \exists! \, p_1 \in P_1 \cap P_2 : [[p_1]] = [[p_2]]$$

Definitie. De reconciliatie van twee entiteiten p_1 en p_2 is de vaststelling dat $[[p_1]]=[[p_2]]$.

Voorbeeld. Figuur 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 2 Conciliatie van twee sets van persoonsentiteiten



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

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 *entiteitgelijkenis* genoemd.

De conciliatie in figuur 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 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 *objectgelijkenis* genoemd.

1.2 Aanleiding

Binnen het WODC speelt de strafrechtsketen – van verdenking tot vervolging en berechting – 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.

Het koppelen van informatiebronnen zonder aanwezigheid van sterke sleutels is de eerste stap naar de bouw van een datawarehouse. In dit datawarehouse worden verschillende informatiebronnen opgenomen, waarbij de gemeenschappelijke entiteittypen zoveel mogelijk worden geconcilieerd. In de eerste versie van het datawarehouse zijn de bronnen HKS, OMDATA en OBJD opgenomen.

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 gelijkenis 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 gelijkenis op verschillende plaatsen in het informatiemodel bijdraagt aan de gelijkenis van één centraal gemeenschappelijk entiteittype: de objectgelijkenis. De objectgelijkenis wordt vervolgens gebruikt in de reconciliatie van objectgelijke entiteiten. De theorie wordt getoetst door middel van een casus. Er wordt 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 hoofdstuk 2 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 Gegevensgebied

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.

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.

2.1.1 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 kun-

nen worden, is het in de praktijk zo dat er meerdere processen-verbaal onder een zaak geregistreerd kunnen staan.

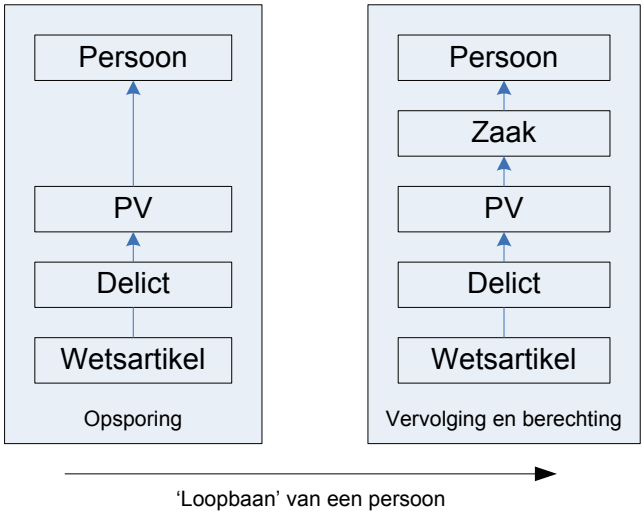
Delict
Een delict is een misdrijf gepleegd door een verdachte. Delicten worden ook wel *strafbare feiten* (kortweg feiten) genoemd. Een delict wordt beschreven door één of meer wetsartikelen. Verder worden delicten ingedeeld in rubrieken en subrubrieken volgens de CBS-standaardclassificatie.

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.

2.1.2 Overzicht

De relaties tussen bovengenoemde objecten zijn, gezien vanuit de strafrechtsketen, weergegeven in figuur 3.

Figuur 3 Relatie tussen de entiteittypen in de strafrechtsketen



In de loop van de strafrechtsketen komt het entiteittype Zaak erbij: een procesverbaal 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 (zie bijlage 1) 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 gebruikt, waarvan ook de karakteristieken worden getoond.

Tabel 1 Karakteristieken HKS

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

Tabel 2 Karakteristieken OMDATA

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

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 (zie bijlage 1). 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 schema-integratie zijn twee gebieden te onderscheiden: schema-integratie *met* gemeenschappelijke sleutels (Agarwal et al., 1995; Kim et al., 1993; DeMichiel, 1989; Lim, Srivastava & Prakhakar, 1993) en *zonder* gemeenschappelijke sleutels (Dey et al., 2002; Wang & Madnick, 1989). 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* (Dey et al., 1998), *instance identification* (Wang & Madnick, 1989), *merge/purge problem* (Gass, 1986), *object isomerism* (Chen et al., 1996), *common identifier problem* (Hérnandez & Stolfo, 1995)); in dit onderzoek wordt de term entiteitreconciliatie gebruikt (Dey et al., 2002) 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 gelijkenis tussen twee entiteiten te bepalen. Dit kan automatisch gebeuren, op basis van definitie (Czejdo et al., 1987; Breitbart et al., 2003; Templeton et al., 1987) of op basis van inhoud (Kim et al., 1993; DeMichiel, 1989). Het kan ook gebeuren door bronexperts of door een combinatie van beide (Dey et al., 2002). Bij automatische attribuutselectie spelen vele

problemen een rol (Dey et al., 2002; Cohen, 1998; Monge & Elkan, 1996), waar dit onderzoek zich niet op concentreert. Voor de beschrijving van gelijkenis wordt gebruik gemaakt van expertkennis. De complexiteit van attribuutselectie speelt daarom niet in dit onderzoek: de attribuutselectie is onderdeel van de beschrijving van gelijkenis.

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. (2004, 2006) bespreken hoe deze informatie gebruikt kan worden in relationele informatiebronnen.

Het identificeren van entiteiten kan een probleem zijn (Lim et al., 1993; Prabhakar et al., 1993). 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 gelijkenis gemodelleerd kan worden rond het gemeenschappelijke entiteittype dat geconcilieerd wordt. In de derde paragraaf wordt beschreven hoe uit gelijkenis op attribuutniveau de objectgelijkenis tussen twee entiteiten van het gemeenschappelijk entiteittype wordt berekend. De reconciliatie van objectgelijke entiteiten wordt besproken in de vierde paragraaf. In paragraaf 3.5 is de theorie formeel uitgewerkt. Paragraaf 3.6 geeft een korte bespreking van de theorie, gevolgd door een conclusie.

3.1 Beschrijving van gelijkenis

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

3.1.1 Definitie van gelijkenis

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). Gelijkenis gaat hierin nog een stap verder, en staat ook uitkomsten tussen 0 en 1 toe. 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 objectgelijkenis. De regels om objectgelijkenis te beschrijven worden vastgelegd in een definitie.

Definitie. Een gelijkeniswaarde beschrijft de mate van *object*gelijkenis 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 objectgelijkenis door onwetendheid niet bepaald kan worden.

Een gelijkeniswaarde van 0 betekent ongelijkheid; de twee vergeleken elementen kunnen niet bij hetzelfde object horen. Een gelijkeniswaarde 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 objectgelijkenis tussen twee elementen te bepalen, bijvoorbeeld als één van de elementen informatie mist.

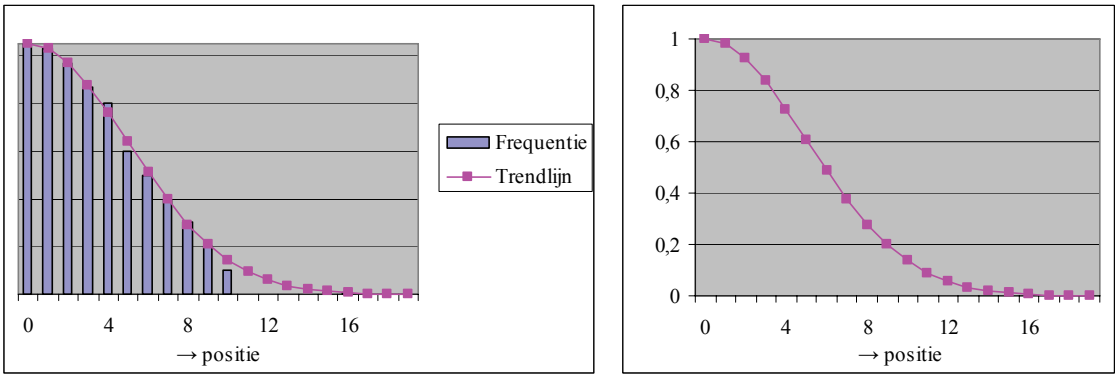
3.1.2 Gemeenschappelijke eigenschappen

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 representatieverschillen, etc. – een gelijkeniswaarde 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 onderlinge positionering (kortweg positie).

Afhankelijk van de attributen kan de positie op verschillende manieren berekend worden (Dey et al., 2002). 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 positie 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 positie berekend worden. Deze positie 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 positieverdeling. Een positie in de positieverdeling wordt vervolgens door middel van een gelijkenisfunctie omgezet in een gelijkeniswaarde.

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 4 toont de verwachte positieverdeling (links) en de bijbehorende gelijkenisfunctie (rechts).

Figuur 4 Positieverdeling en frequentiefunctie (l) en gelijkheidsfunctie (r)



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

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

Definitie. De gelijkenisfunctie $sim(a_1, a_2)$ berekent de gelijkeniswaarde van een gemeenschappelijke eigenschap en wordt gedefinieerd als:

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

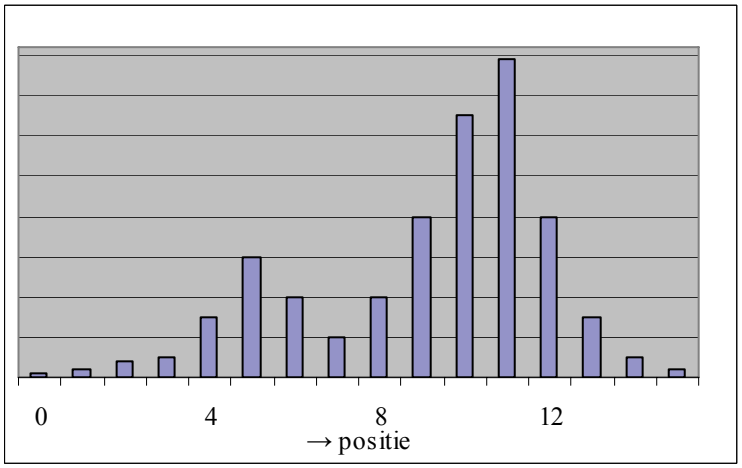
Doordat de frequentie gedeeld wordt door het maximum van de frequentiefunctie, heeft de gelijkenisfunctie een bereik van 0 tot en met 1. Als de positie $\delta(a_1, a_2)$ niet bepaald kan worden, dan heeft de gelijkenisfunctie als uitkomst *n.a.*

Voor sterke eigenschappen levert de gelijkenisfunctie 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 positieverdelingen
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 gelijkenisfuncties te definiëren voor elke gemeenschappelijke eigenschap, waarbij de domeinen van de gelijkenisfuncties 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 positie tussen pleegdatum en eindvonnis is voor snelrechtzaken veel kleiner dan voor andere zaken. Stel dat het histogram er als volgt uit ziet (de posities zijn weken):

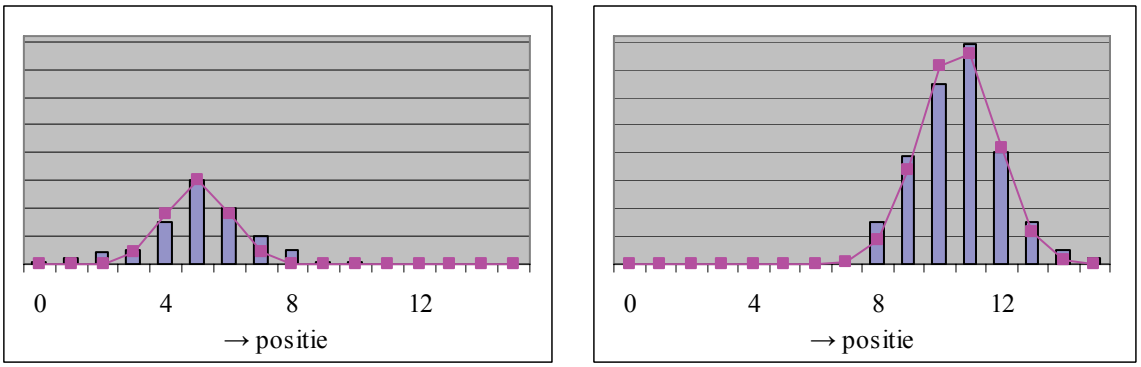
Figuur 5 Positieverdeling voor snelrecht en overige zaken samen



Het histogram heeft twee lokale maxima op de posities 5 en 11; voor elke gemeenschappelijke eigenschap één. Door de eigenschappen te scheiden, kan voor allebei een positieverdeling worden vastgesteld. Elke positieverdeling heeft

zijn eigen gelijkenisfunctie, waardoor de piek in figuur 6 (links) nu ook maximale gelijkenis oplevert.

Figuur 6 Positieverdeling en frequentiefunctie voor snelrecht (links, d_0=5) en voor overige zaken (rechts, d_0=11)



Ruis
In sommige gevallen kan de positie 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 gelijkenis nagenoeg nul is. Als het belangrijk is om de ruis op te vangen, dan moet de gelijkenis voor deze gevallen verhoogd worden. Ruis is echter niet te onderscheiden van reële posities. 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 gelijkenis oplevert. De ruisdrempel moet daarom alleen worden ingezet voor de posities waar de ruis verwacht wordt.

Voorbeeld. Gegeven het voorbeeld in figuur 4. 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 positieverdeling 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 gelijkenis bij een positie 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 gelijkenis tussen twee attributen berekend kan worden en levert gelijkeniswaarde op. Aangezien een attribuut meerdere eigenschappen kan beschrijven, kan een kennisregel opgebouwd zijn uit meerdere gelijkenisfuncties.

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 gelijkenis 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 gelijkenis. 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:

Tabel 3 Sets van persoonsentiteiten

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

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

Tabel 4 Mogelijkheden per attribuut

geb.datum	mogelijkheden	geslacht	mogelijkheden
15-02-1965	1	m	9
17-07-1974	1	v	9
01-01-1975	4		
23-04-1977	1		
01-03-1980	1		

Duidelijk is dat geboortedatum beter in staat is het aantal mogelijkheden te beperken dan geslacht. De kennisregel die de gemeenschappelijke 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 (zie bijlage 1) is dit verder uitgewerkt.

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

3.2 Modellering van gelijkenis

Deze paragraaf bespreekt de distributie van gelijkenis. 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[2] en zullen daarom meestal behoren tot een gemeenschappelijk entiteittype. Als de attributen geen gemeenschappelijk entiteittype hebben of verschillende entiteittypen hebben[3], 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:

Tabel 5 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

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

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 gelijkenis

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 gelijkeniswaarden uit attribuutvergelijkingen bij de centrumknoop komen. Het doorgeven ('distributie') van gelijkenis 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 gelijkenis beschrijft, wordt de *gelijkenisdistributie* genoemd. Later in deze paragraaf wordt een exacte definitie van dit begrip gegeven.

De gelijkenisdistributie 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 gelijkenis van de centrumknoop beschrijft, is het wenselijk elke knoop zo goed mogelijk te beschrijven. In de gelijkenisdistributie worden gerichte lijnen geplaatst tussen knopen met een directe relatie.

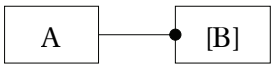
Voorbeeld. Stel, er is een gelijkenisdistributie $G(V,E,c)$ met $V = \{A,B,C\}$, $E = \{AC,BC\}$ en $c = C$. Stel nu dat de relatie AB ook bestaat. Dan is de opname van de relatie AC gebaseerd op de transitieve afsluiting $AB \wedge BC \rightarrow AC$. In dit geval moet niet AC, maar AB worden opgenomen: $E' = \{AB,BC\}$. 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 gelijkenis van delicten door te geven via hun proces-verbaal, wordt het proces-verbaal ook beter beschreven.

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

Definitie. Een gelijkenisdistributie $G(V,E,c)$ 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 gelijkenisdistributie betekent dat de gelijkenis van een knoop naar een andere knoop moet worden doorgegeven. De gelijkenis in een knoop kan pas berekend worden als de gerelateerde knopen achter alle inkomende pijlen zijn berekend, d.w.z. de gelijkenis 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 7 Knoop A distribueert gelijkenis naar centrumknoop B



Binnen de gelijkenisdistributie worden twee soorten gelijkenis gedefinieerd.

Definitie. De entiteitgelijkenis s_A beschrijft de gelijkenis binnen een knoop A.
Definitie. De objectgelijkenis S(A) is de formule voor de totale gelijkenis in A.

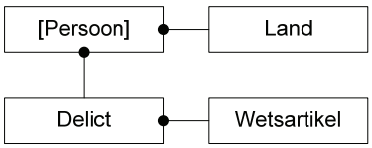
De formule voor objectgelijkenis S(A) bestaat uit de entiteitgelijkenis van A en de objectgelijkenis van knopen achter inkomende pijlen naar A. Binnen de formule S(A) wordt de operator $\oplus$ gebruikt, die de verschillende gelijkeniswaarden combineert. De operator en entiteitgelijkenis wordt uitgewerkt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

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

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

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

Figuur 8 Voorbeeld van een gelijkenisdistributie



Voorbeeld. Gegeven de gelijkenisdistributie uit figuur 8. In stap 1 worden Land en Wetsartikel berekend. In stap 2 wordt de objectgelijkenis van Land en Wetsartikel gedistribueerd naar respectievelijk Persoon en Delict. In stap 3 wordt Delict berekend. Vervolgens wordt stap 2 herhaald door de objectgelijkenis van Delict te distribueren naar Persoon en wordt in stap 3 de objectgelijkenis van Persoon berekend.

De knopen in het voorbeeld worden in een bepaalde volgorde berekend, die afhankelijk is van de gelijkenisdistributie. Als laatste wordt de objectgelijkenis van de centrumknoop berekend. De volgorde waarin objectgelijkenis wordt berekend, wordt weergegeven in een gelijkenisstelsel.

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

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

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

Door distributie van gelijkenis van boven naar beneden geldt uiteindelijk $S(P)=s_L \oplus s_P \oplus s_D \oplus s_W$.

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

Een cykel in de gelijkenisdistributie 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 gelijkenisdistributie $G=(V,E,c)$ met $V=\{A,B\}$, $E=\{AB,BA\}$ en $c=B$. Er geldt: $S(A)=s_A \oplus S(B)$ en $S(B)=s_B \oplus S(A)$. Dit stelsel is onoplosbaar. Door eerst A te berekenen met alleen de entiteitgelijkenis uit B is het stelsel wel oplosbaar: $S(A)=s_A \oplus s_B$ en $S(B)=s_B \oplus S(A)$.

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 gelijkenisberekening op te nemen. In een ternaire relatie wordt elke knoop beschreven door twee andere knopen. Er geldt: $S(A)=s_A \oplus S(B) \oplus S(C)$, $S(B)=S(A) \oplus s_B \oplus S(C)$ en $S(C)=S(A) \oplus S(B) \oplus s_C$. 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 gelijkenisstelsel zoals beschreven in de vorige paragraaf wordt de gelijkenis 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 gelijkenisdistributie uit figuur 8, 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 delictcombinaties te berekenen waarvoor de persooncombinatie (tot dan toe) niet ongelijk is.

De distributie van gelijkenis 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 6 toont mogelijke vergelijkingen. Hiervoor worden de formules voor objectgelijkenis uitgebreid met entiteitidentificaties. Zo levert de formule $S(L)(1,2)$ de objectgelijkenis tussen landsentiteiten uit bron I en II met respectievelijk een identificatie van 1 en 2. De entiteitcombinaties en hun gelijkeniswaarden zijn voorbeelden.

Tabel 6 Verschillende combinaties

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)
...	...	...

Persoonscombinaties worden alleen berekend als de landcombinatie niet ongelijk is. Evenzo worden delictcombinaties alleen berekend als de persoons-

combinatie berekend én niet ongelijk is. Het aantal vergelijkingen wordt hiermee verlaagd. Hoe meer vergelijkingen in een begin stadium ongelijkheid opleveren, hoe efficiënter de gelijkenisberekening. 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 (zie bijlage 1). 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 gelijkenisstelsel van de gelijkenisdistributie: $S(L)=s_L$, $S(D)=s_D$ en $S(P)=S(L) \oplus s_P \oplus S(D)$. 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 $s_P \oplus S(D)$ 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 objectgelijkenis in een centrumknoop (c) bestaat in het algemeen uit maximaal één ouderknoop (o), de entiteitgelijkenis en kindknopen ($k_1..k_n$). De ouderknoop kan op zijn beurt weer een ouderknoop (p) en kindnopen ($l_1..l_m$) hebben. Eén van de kinderen is de centrumknoop, stel $l_m=c$. Eerst wordt een eenvoudig geval besproken met $m=2$ en $n=1$. Het gelijkenisstelsel 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 gelijkenisstelsel 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 entiteitgelijkenis 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 gelijkenisstelsel S wordt nu:

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

Een formule voor objectgelijkenis wordt grafisch weergegeven door middel van een informatiemodel.
Definitie. Een informatiemodel is een hiërarchisch datamodel van de knopen in een gelijkenisdistributie met hun ouder/kind relaties, waarmee een formule voor objectgelijkenis wordt weergegeven. In het informatiemodel worden soms ook de gebruikte attributen weergegeven.

Het informatiemodel is een *hiërarchisch datamodel* (Silberschatz et al., 2005) 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 gelijkenisdistributie $G(V,E,c)$ met $V=\{A,B,C,D,E,F,G,H\}$, $E=\{AC,BA,DA,EB,FC,GC,HD\}$ en $c=C$. De gelijkenisdistributie en clustering wordt beschreven door $\underline{c} = a(b(e)d(h))(c(fg))$.

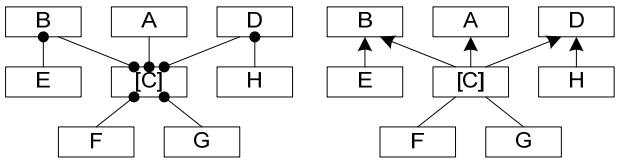
Figuur 9 Gelijkenisdistributie (l) en bijbehorend informatiemodel (r)

De gelijkenisdistributie wordt in dezelfde boomstructuur weergegeven als het informatiemodel. Op die manier kan aan de structuur van de gelijkenisdistributie het bijbehorende informatiemodel worden afgelezen (zie figuur 9).

Uit voorgaande tekst blijkt dat de gewenste clustering alleen mogelijk is in een hiërarchische boomstructuur. Met andere woorden: de gelijkenisdistributie kan alleen gebruik maken van clustering als het bijbehorende informatiemodel een hiërarchisch datamodel is. Het hiërarchisch datamodel legt twee beperkingen op aan de gelijkenisdistributie:
— meer-op-meer relaties zijn niet toegestaan
— een knoop mag slechts één ouderknoop hebben

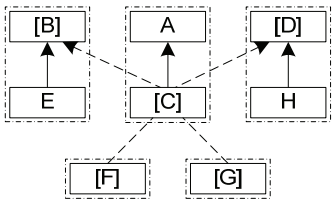
Voorbeeld. Gegeven een gelijkenisdistributie en het bijbehorende datamodel, met meer-op-meer relaties en een centrumknoop met meerdere ouderknopen:

Figuur 10 Gelijkenisdistributie (l) en bijbehorend datamodel (r)



In deze gelijkenisdistributie 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 11. Hierin zijn de informatiemodellen omrand. De virtuele relaties tussen knopen in verschillende modellen zijn gestippeld.

Figuur 11 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 gelijkeniswaarden op het gewenste moment berekend. Bovendien wordt zoveel mogelijk gebruik gemaakt van clustering. Er ontstaat een gelijkenisstelsel, waarin elke formule de objectgelijkenis van een centrumknoop in een hiërarchisch model berekent. Bij figuur 11 hoort bijvoorbeeld het volgende gelijkenisstelsel *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 11 vormen een link tussen de verschillende informatiemodellen en geven aan dat de gelijkenis 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 (Silberschatz, 2005). Virtuele relaties bevatten ook slechts een verwijzing. Deze verwijzing wordt gebruikt om de gelijkenis uit het andere informatiemodel op te halen.

Virtuele relaties – zoals in het gelijkenisstelsel *S* bij figuur 11 – worden voor de knoop geplaatst waarin ze nodig zijn, zonder gebruik van haakjes: het zijn immers geen ouder- of kindknopen. De gelijkeniswaarden 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 gelijkenis

Deze paragraaf laat zien hoe de gelijkenis wordt berekend in een informatiemodel. Bij het rekenen met gelijkenis wordt gebruik gemaakt van een gelijkenisoperator ($\oplus$). De gelijkenisoperator wordt gebruikt om uit verschillende gelijkeniswaarden één waarde te berekenen. Een gelijkeniswaarde 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 gelijkeniswaarde kan het gedrag van de operator worden afgeleid.

De operator berekent één gelijkeniswaarde uit meerdere gelijkeniswaarden. De gelijkeniswaarden tellen even zwaar: de operator berekent een gemiddelde over de gelijkeniswaarden 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 gelijkeniswaarden a_1, ..., a_n. De n-aire operator $a_1 \oplus ... \oplus a_n$ wordt gedefinieerd als:

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

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

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

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 gelijkeniswaarden a=0.5, b=0.5 en c=1. Er geldt: $(a \oplus b) \oplus c = (1/2) \oplus 1 = 0.5 \oplus 1 = 1.5/2 = 0.75$ en $a \oplus (b \oplus c) = 0.5 \oplus (1.5/2) = 0.5 \oplus 0.75 = 1.25/2 = 0.625$. Dus $(a \oplus b) \oplus c \neq a \oplus (b \oplus c)$, de operator is niet associatief.

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

Definitie. Gegeven n gewogen gelijkeniswaarden $a_1[g_1]$, ..., $a_n[g_n]$. De n-aire operator wordt $a_1[g_1] \oplus ... \oplus a_n[g_n]$ gedefinieerd als:

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

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

3.3.1 Van attribuutgelijkheid naar entiteitgelijkenis

Kennisregels beschrijven de gelijkenis tussen twee gemeenschappelijke attributen. Bij de vergelijking van een entiteitcombinatie heeft een kennisregel één gelijkeniswaarde. Verder heeft elke kennisregel een gewicht (zie paragraaf 3.1.3).

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

Definitie. De entiteitgelijkenis van K wordt gedefinieerd als:

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

De gelijkeniswaarde r_i kan berekend zijn uit meerdere gelijkeniswaarden, 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 de waarde van een persoonsattribuut op die manier onzeker 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 gelijkeniswaarde 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 (zie bijlage 1) worden enkele alternatieven besproken.

3.3.2 Van entiteitgelijkenis naar objectgelijkenis

De entiteitgelijkenis beschrijft de gelijkenis nog niet volledig. Uit de gelijkenisdistributie is bekend dat alle gelijkenis bij de centrumknoop meetelt, maar dat de gelijkenis via andere knopen kunnen worden doorgegeven. Elke knoop kent daarom een vorm van objectgelijkenis. De objectgelijkenis bestaat uit de gelijkenis 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 gelijkeniswaarden wordt de objectgelijkenis berekend:

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

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

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

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

$$S(K) = S(0) \oplus (r_1[w_1] \oplus \ldots \oplus r_m[w_m]) \oplus S(1)$$
$$S(K) = S(0) \oplus r_1[w_1] \oplus \ldots \oplus r_m[w_m] \oplus S(1)$$

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

— hoe meer kennisregels, hoe zwaarder weegt de entiteitgelijkenis;
— hoe zwakker de kennisregels, hoe zwaarder wegen de andere objectgelijkenissen.

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 objectgelijkenis luidt nu:

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

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

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

De formule voor objectgelijkenis bestaat uit de entiteitgelijkenis en objectgelijkenis uit andere knopen. De objectgelijkenis uit andere knopen is één gelijke-

niswaarde die de objectgelijkenis in de formule beschrijft. Zo beschrijft S(B) in S(A)=$s_A \oplus$ S(B) de gelijkenis van entiteitcombinaties uit knoop B die een relatie hebben met de entiteitcombinatie uit knoop A waarvoor de objectgelijkenis berekend wordt. Als knoop B een kindknoop is van knoop A, dan kunnen dit meerdere entiteitcombinaties, en daarmee meerdere gelijkeniswaarden, zijn. Het berekenen van één gelijkeniswaarde uit meerdere gelijkeniswaarden 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 gelijkeniswaarde berekend wordt uit meerdere gelijkeniswaarden.

3.4.1 Selecteren van de reconciliaties

Voor alle entiteitcombinaties is de objectgelijkenis 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. (1998, 2002) maken ook gebruik van positie 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 gelijkenismatrix 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 gelijkeniswaarde van de bijbehorende elementcombinatie.

Definitie. Een gelijkenismatrix *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 gelijkeniswaarde van de bijbehorende entiteitcombinatie. Bij een gelijkeniswaarde *n.a.* is de cel 0, omdat op basis van missende informatie een combinatie nooit gereconcilieerd kan worden.

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

Definitie. Een reconciliatiemogelijkheid is een entiteitcombinatie met een gelijkeniswaarde groter dan de gelijkenisdrempel. De gelijkeniswaarden *n.a.* en 0 leveren nooit een reconciliatiemogelijkheid op.

In de reconciliatiemethode wordt naast de gelijkeniswaarde 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-gelijkenismatrix 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 gelijkeniswaarden 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 gelijkeniswaarde en waarschijnlijkheid worden gecombineerd in één waarde: de reconciliatiescore. Hierin heeft de gelijkeniswaarde een bepaald gewicht ten opzichte van de waarschijnlijkheid. Dit gewicht wordt aangegeven door de gelijkenisfactor *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 gelijkeniswaarde en waarschijnlijkheid. In dit onderzoek is de

gelijkenisfactor geoptimaliseerd voor de beste resultaten. Uiteindelijk is een gelijkenisfactor 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 gelijkeniswaarde

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 gelijkeniswaarde 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 gelijkenis van een set van element-combinaties K samen in één gelijkeniswaarde.

Definitie. De samenvoegfunctie voor gereconcilieerde gelijkenis $m(K)$ 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 (zie bijlage 1).

3.5 Formele theorie

In deze paragraaf is de voorgaande theorie formeel uitgewerkt. Indien van toepassing is verwezen naar verwante definities en andere passages in de eerdere tekst.

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) $\forall a \in A, o \in O : o \in \mathrm{dom}\, a$ ($a(o)$ 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:

> $\forall e \in E : e \in \mathrm{dom}\, a$ ($a(e)$ is zinvol)

Opmerking. Zie intuïtie 1 op pagina 37 voor de mogelijkheid dat $e.a \neq [[e]].a$.

Verwijzing. Buiten deze paragraaf wordt gesproken over 'entiteittype' in plaats van de bredere definitie 'entiteitsoort', omdat in dit onderzoek de aanname gedaan wordt dat een reconciliatie slechts tussen twee entiteiten plaatsvindt (één-op-één koppeling). Dit impliceert over het algemeen éénzelfde entiteittype. Toekomstig onderzoek zou ook meer-op-meer koppelingen toe kunnen staan, waarbij de definitie entiteitsoort ingezet kan worden.

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 16, 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.

Opmerking. Het begrip 'positie' dient niet verward te worden met 'afstand', aangezien '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 $e.a \neq o.a$, dan verwachten we "*positie*(*e.a*, *o.a*) is klein".

Definitie 7. De mate van overeenkomst drukken we uit in een gelijkeniswaarde:

$$[0..1] \cup \{"n.a."\}$$

Verwijzing. Deze definitie komt overeen met de definitie van gelijkeniswaarde op pagina 17.

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:

a entiteit, object: $[[e]] = o \equiv e.s = o.s$

b 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:

$$freq(d) = \frac{\#\{o \in O \mid positie_{\vec{a}_1,\vec{a}_2}(\vec{a}_1,\vec{a}_2) = d}{\#(\{o \in O\}}, \text{met } d \in \text{ran } positie_{\vec{a}_1,\vec{a}_2} - \{"n.a."\}$$

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 18, 2e definitie).

Definitie 10. Laat p = ($\vec{a}_1$, $\vec{a}_2$, *freq*) een gemeenschappelijke eigenschap voor E_1, E_2 zijn. De gelijkenis in de gemeenschappelijke eigenschap p

$$sim_p : E_1 \times E_2 \rightarrow \text{gelijkeniswaarde}$$

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 gelijkenis in een gemeenschappelijke eigenschap een gelijkenisfunctie genoemd (definitie op pagina 19).

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:

$$\forall e_1 : E_1, e_2 : E_2 \forall i, j \in \{1,..,n\}, i \neq j \bullet \neg \quad ((e_1.\vec{a}_{i1} \in D_{i1} \wedge e_2.\vec{a}_{i2} \in D_{i2}) \wedge (e_1.\vec{a}_{i1} \in D_{i1} \wedge e_2.\vec{a}_{i2} \in D_{i2}))$$

$$\left(\langle D_{i1} \times D_{i2} \rangle \text{ is een (deel)partitionering van ran}(\vec{a}_{i1}, \vec{a}_{i2})\right)$$

De gelijkenis in de disjuncte gemeenschappelijke eigenschappen $p_1..p_n$

$$sim_{(p_1,..,p_n)} : E_1 \times E_2 \rightarrow \text{gelijkeniswaarde}$$

wordt gedefinieerd als:

$$sim_{(p_1,..,p_n)}(e_1, e_2) = \begin{cases} sim_{p_1}(e_1, e_2) & \text{als } e_1.\vec{a}_{11} \in D_{11} \wedge e_2.\vec{a}_{12} \in D_{12} \\ \vdots & \vdots \\ sim_{p_n}(e_1, e_2) & \text{als } e_1.\vec{a}_{n1} \in D_{n1} \wedge e_2.\vec{a}_{n2} \in D_{n2} \\ \text{"}n.a.\text{"} & \text{anders} \end{cases}$$

We noemen $sim_{(p_1,..,p_n)}$ een A-kennisregel voor E_1, E_2.

Verwijzing. In hoofdstuk 3 (definitie op pagina 19) 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 $x_1..x_n$ gelijkeniswaarden zijn (dus in $[0..1] \cup \{\text{"}n.a.\text{"}\}$). De n-aire gelijkenisoperator wordt gedefinieerd als:

$$x_1 \oplus ... \oplus x_n = \begin{cases} 0 & \text{als } \exists i : x_i = 0 \\ \text{"}n.a.\text{"} & \text{als } \forall i : x_i = \text{"}n.a.\text{"} \\ \dfrac{\sum_{i:x_i \neq \text{"}n.a.\text{"}} x_i}{\#\{i \mid x_i \neq \text{"}n.a.\text{"}\}} & \text{anders} \end{cases}, \text{ met } i \in \{1,..,n\}$$

Opmerking. Het symbool $\oplus$ is suggestief. De gelijkenisoperator heeft echter zowel eigenschappen van een optelling, als van een vermenigvuldiging. Het symbool $\otimes$ is daarom ook een alternatief.

Verwijzing. Deze definitie komt overeen met de definitie van gelijkenisoperator zonder gewicht op pagina 30. De stelling op pagina 30 toont aan dat de gelijkenisoperator 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 gelijkenis tussen entiteiten (entiteitgelijkenis)

$$s_O : E_1 \times E_2 \rightarrow \text{ gelijkeniswaarde}$$

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). Volledige zou zijn om het stel kennisregels als parameter mee te geven, bijvoorbeeld als subscript aan "*s*".

Verwijzing. Deze definitie komt overeen met de definitie van entiteitgelijkenis op pagina 30 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 aan het einde van paragraaf 3.2.1.

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: $K_1 \to K_2 \equiv (\exists p : K_1 \xrightarrow{P} K_2)$.

3 $wortel \in Kn$, zdd: $\forall k \in Kn : k \sim^* wortel$

4 Er zijn geen cykels: $\forall k \in Kn : \neg(k \sim^+ k)$

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 gelijkenisdistributie op V als $Kn \subseteq Es \wedge Pn \subseteq Rs$.

Verwijzing. Deze definitie komt overeen met de definitie van gelijkenisdistributie op pagina 23.

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 gelijkenis-distributie 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.

De gelijkenis tussen entiteiten en hun gerelateerde entiteiten (objectgelijkenis)

$$S_K : E_1 \times E_2 \to \text{gelijkeniswaarde}$$

wordt gedefinieerd als

$$\begin{aligned} S_K(e_1, e_2) : K = s_O(e_1, e_2) \quad & \oplus\ T_P(S_{K' \xrightarrow{P} K}(e_1, e_2)) \\ & \vdots \\ & \oplus \text{ voor alle } P \in Pn \text{ met } doel(P) = K \end{aligned}$$

waarbij:

— $S_{K, K'}$ de gelijkenisbijdrage van entiteiten uit *K'* aan de objectgelijkenis van entiteiten uit *K* is (zie definitie 19);

— T_P gepostuleerd wordt voor alle $P \in Pn$;

— in de berekening voor $s_O(e_1, e_2)$ de vergelijking zelf wordt ingevuld – in plaats van de uitkomst – voordat de gelijkenisoperator wordt uitgevoerd.

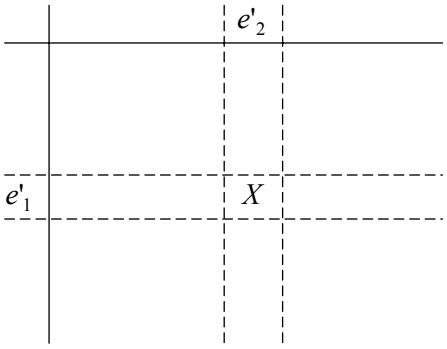
Verwijzing. De definitie voor objectgelijkenis is gelijk aan de definitie van objectgelijkenis op pagina 24 zonder gewichten voor kennisregels. De transformatiefunctie T_p wordt besproken in paragraaf 3.4.2.

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 gelijkenisbijdrage van entiteiten uit *K'* aan de objectgelijkenis van entiteiten uit K

$$S_{K' \xrightarrow{P} K} : E_1 \times E_2 \to \text{matrix van gelijkeniswaarden}$$

wordt gedefinieerd als:

$S_{K' \xrightarrow{P} K}(e_1, e_2) : K =$



met $X = S_{K'}(e'_1, e'_2) : K'$, $(e'_1, e_1) \in R_1$ 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.

Laat $K = (E_1, E_2) \in Kn$ een knoop uit V, G zijn, en $e_1 \in E_1$, $e_2 \in E_2$.

Definitie 20. De *definitieve* objectgelijkenis tussen entiteiten

$$S'_K : E_1 \times E_2 \to [0..1]$$

wordt gedefinieerd als

$$S'_K(e_1, e_2) = \begin{cases} S_K(e_1, e_2) & \text{als } S_K(e_1, e_2) \neq \text{"}n.a.\text{"} \\ 0 & \text{anders} \end{cases}$$

Definitie 21. 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

$$prob_K(e_1, e_2) = \frac{S'_K(e_1, e_2)}{\sum_{e \in E_2} S'_K(e_1, e)} \cdot \frac{S'_K(e_1, e_2)}{\sum_{e \in E_1} S'_K(e, e_2)}$$

Verwijzing. Deze definitie is gelijk aan de definitie van waarschijnlijkheid op pagina 34.

Intuïtie 2. De kans dat een paar entiteiten (e_1, e_2) hetzelfde object representeren, neemt *toe* als:
— de mate van overeenkomst *toe*neemt (*gelijkenis*)
— de mate van overeenkomst in gerelateerde entiteitparen *af*neemt (*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: Laat $a_K = S'_K(e_1, e_2)$ en $pr_K = prob_K(e_1, e_2)$. Dan is de intuïtie dat geldt:

$$P(e'_2 \in E_2 \mid a_K >> S'_K(e_1, e'_2) \vee (a_K \approx S'_K(e_1, e'_2) \wedge pr_K \geq prob_K(e_1, e'_2)) \bullet [[e_1]] = [[e_2]])$$
$$\geq$$
$$P(e'_2 \in E_2 \mid a_K << S'_K(e_1, e'_2) \vee (a_K \approx S'_K(e_1, e'_2) \wedge pr_K \leq prob_K(e_1, e'_2)) \bullet [[e_1]] = [[e_2]])$$

Opmerking. Doordat objecten binnen één entiteitsoort slechts door één entiteit gerepresenteerd kunnen worden (zie definitie 6), geldt in bovenstaande vergelijking:

$$[[e_1]] = [[e_2]] \Rightarrow ([[e_1]] \neq [[e'_2]] \vee e_2 = e'_2)$$

De operatoren >> en << betekenen respectievelijk “veel groter dan” en “veel kleiner dan”. Er geldt in de vergelijking:

$$(a_K >> S'_K(e_1, e'_2) \vee a_K << S'_K(e_1, e'_2)) \Leftrightarrow \neg(a_K \approx S'_K(e_1, e'_2))$$

Verwijzing. In deze intuïtie speelt de waarschijnlijkheid alleen een rol als de gelijkeniswaarden dicht bij elkaar liggen. In hoofdstuk 3 zijn gelijkenis en waarschijnlijkheid gecombineerd in één waarde; de reconciliatiescore (zie definitie op pagina 34). 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 *toe*name van het aantal (gebruikte) gemeenschappelijke eigenschappen.

Intuïtie 2 en 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, besproken in hoofdstuk 6 en 7.

3.6 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 positie. Hierdoor kan ook een mate van gelijkenis 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 gelijkenis 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 entiteitgelijkenis berekend. De entiteitgelijkenis uit de knopen wordt gedistribueerd naar de centrumknoop en telt mee in de objectgelijkenis. De gebruiker bepaalt hiervoor een gelijkenisdistributie, die feitelijk berust op de natuurlijke relaties die knopen hebben met de centrumknoop.

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

Expertkennis beschrijft gelijkenis op attribuutniveau. Deze gelijkenis wordt binnen een knoop omgezet in een entiteitgelijkenis. Alle entiteitgelijkenis in het informatiemodel bepaalt uiteindelijk de objectgelijkenis. Aan de hand van de berekende objectgelijkenis 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. In bijlage A (zie bijlage 1) worden deze bronnen nader toegelicht. 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 gelijkenis in een informatiemodel. Tot slot wordt in de laatste paragraaf ingegaan op de gebruikte gegevens.

4.1 Beschrijving van gelijkenis

De expertkennis om gelijkenis te beschrijven is met standaardtechnieken (Durkin, 1994; Scott et al., 1991; Stefik, 1995), waaronder interviews, van bronexperts verkregen. Een uitgebreide beschrijving hiervan is terug te vinden in de documentbijlage Onderzoeksomgeving (zie bijlage 1). Dit heeft de volgende overlap opgeleverd:

Tabel 7 Gevonden overlap

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

De gevonden overlap moeten geformaliseerd worden om de gelijkenis te kunnen beschrijven. Na de formalisatie kunnen de eigenschappen eenvoudig worden omgezet in kennisregels (meer hierover in bijlage E, zie bijlage 1). 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 positieverdelingen zijn er?
- Voor elke positieverdeling:
  - Hoe wordt de positie bepaald?
  - Welke formule beschrijft de verdeling?
  - Wat is het maximum van de formule?
  - Over welk subdomein van de attribuutwaarden gaat de positieverdeling? (alleen bij meerdere positieverdelingen)
  - 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*).

4.1.1 Geboortedatum

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

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

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:

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

4.1.2 Geslacht

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

4.1.3 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 gelijkenisfunctie sim_{land} verwijst naar het informatiemodel voor de reconciliatie van landen. Dit model wordt besproken in bijlage D (zie bijlage 1).

4.1.4 Pleegdatum

Deze kennisregel is het voorbeeld waarin gelijkenis 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 positie tussen deze twee attributen wordt bepaald door de positiefunctie:

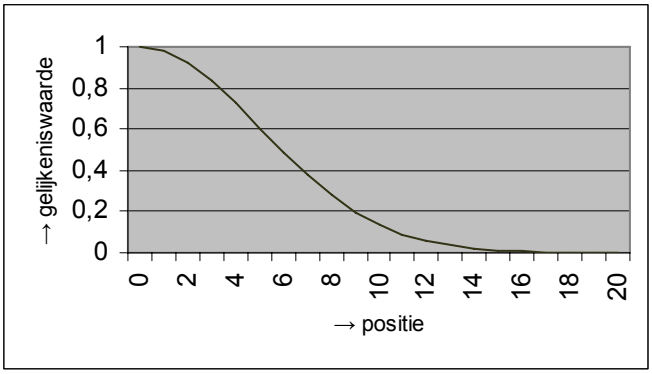
$$\delta = datum_{pv} - datum_{delict}$$

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 gelijkenisfunctie beschrijft de trendlijn die volgens de bronexperts het histogram van de verwachte posities 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 gelijkenisfunctie grafisch weer.

Figuur 12 Gelijkenisfunctie kennisregel 'pleegdatum'



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

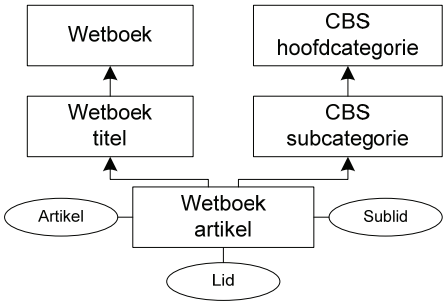
Optimalisatie
Slechts een klein deel van de delicten wordt laat bij de politie aangemeld. Het gaat hier waarschijnlijk ook om een bepaalde categorie zaken (zoals zwaardere

vermogensdelicten). Dit is nog niet in detail onderzocht. Feit is wel dat bij zeer grote posities, 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 maximumpositie ingesteld tot waar de ruisdrempel wordt ingezet. Deze positie is gesteld op 1 jaar.

4.1.5 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 13.

Figuur 13 Schematische weergave indeling wetsartikel (in E/R diagram notatie)

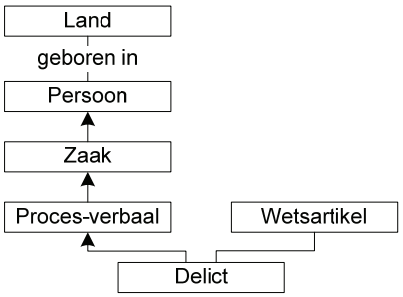


Meer over de categorisering en reconciliatie van wetsartikelen is te vinden in bijlage D (zie bijlage 1).

4.2 Modellering van gelijkenis

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

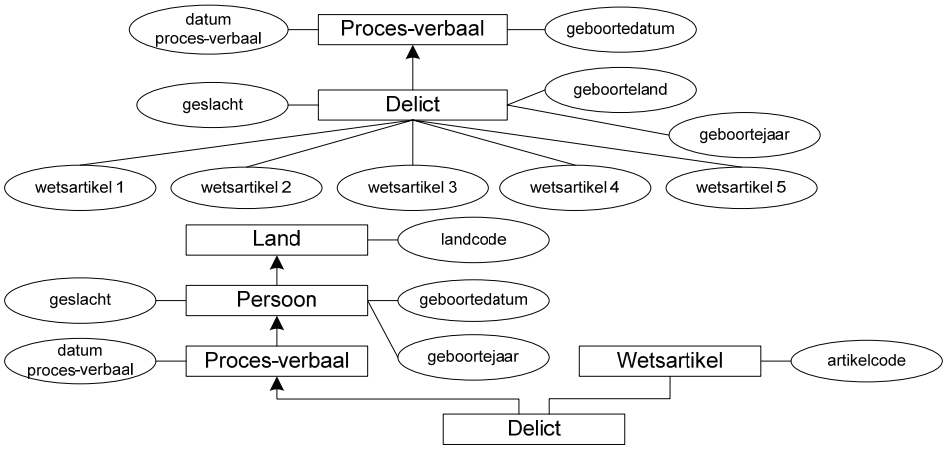
Figuur 14 Gemeenschappelijk datamodel



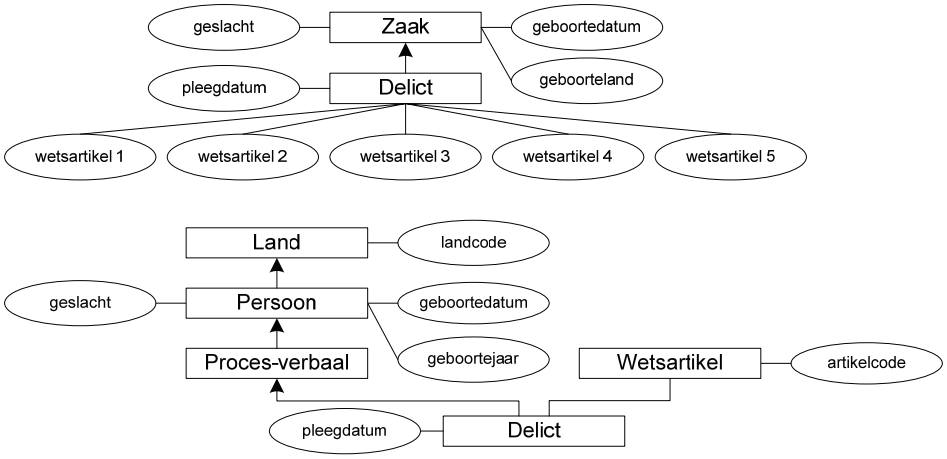
Het datamodel wordt gevormd door de gemeenschappelijke entiteittypen waartoe de attributen van kennisregels behoren. Om de gemeenschappelijke entiteittypen te kunnen bepalen, is het van belang om de ligging van attributen in

elke bron afzonderlijk te kennen, zowel fysiek als logisch (bij het bijbehorende entiteittype).

Figuur 15 HKS fysiek (boven) en logisch (onder)



Figuur 16 OMDATA fysiek (boven) en logisch (onder)



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

4.2.1 Informatiemodel

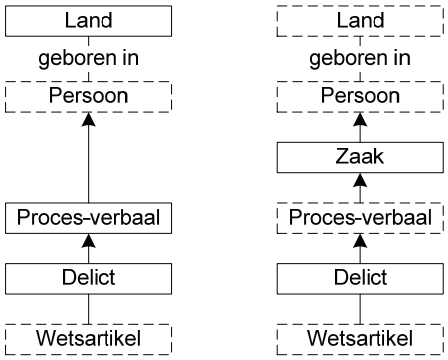
Om de conciliatie van twee bronnen te kunnen uitvoeren, moeten de attributen van elke kennisregel worden geplaatst onder een gemeenschappelijk entiteittype. In figuur 15 en figuur 16 is te zien dat dit alleen bij de kennisregel ‘pleegdatum’ niet het geval is. Ook sommige entiteittypen moeten nog worden 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.

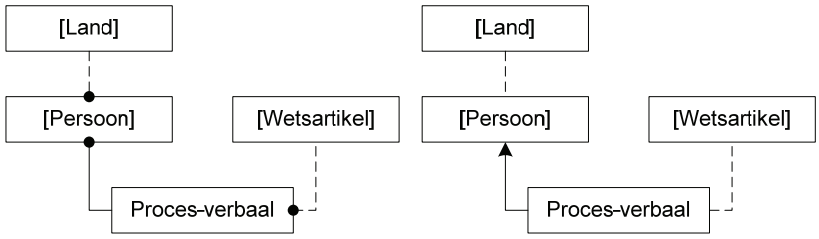
Ook bij wetsartikelen is geen sprake van een vast gegeven. Naarmate het strafproces vordert kunnen er wetsartikelen veranderen: soms komen er nieuwe bewijzen, waardoor een ander wetsartikel het delict beter beschrijft of de aanklacht versterkt kan worden. Reconciliatie van wetsartikelen onder een delict 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 17 Datamodellen van HKS en OMDATA met afgeleide entiteittypen



Figuur 18 Gelijkenisdistributie (l) en informatiemodellen (r)



Het behorende gelijkenisstelsel S is:

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

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 (zie bijlage 1).

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.

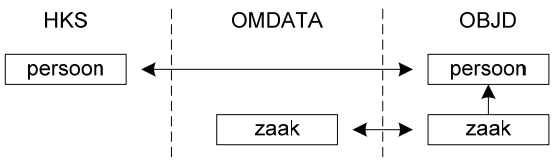
Voor de referentieset is extra persoonsinformatie gebruikt om de persoonsentiteiten handmatig te reconciliëren. Deze informatie, maar ook het proces zelf, kan daarom niet beschreven worden in dit document. De referentieset is met de grootst mogelijke nauwkeurigheid samengesteld. Daarom worden de correcte 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 19 Koppelingen in de referentieset

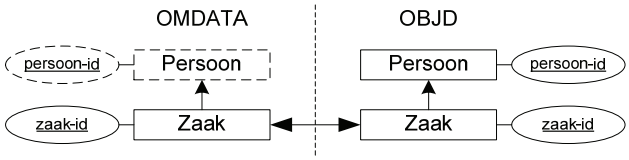


Om het persoonsniveau vanuit OBJD in OMDATA te plaatsen, zijn uit de referentieset de OMDATA-zaken geleverd, uitgebreid met een uniek geanonimiseerd persoonsnummer (een soort volgnummer). Voor de toetsing zijn verder voor de gereconcilieerde persoonsentiteiten de bijbehorende metanummers uit HKS geleverd.

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 20 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 reconciliatiemogelijkheden en het maken van reconciliaties;
— het analyseren van de kwaliteit van de conciliatie.

De theorie bespreekt 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 gelijkenisdistributie – moeten keuzes gemaakt worden in de richting waarin gelijkenis 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 reconciliatiemogelijkheden 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.

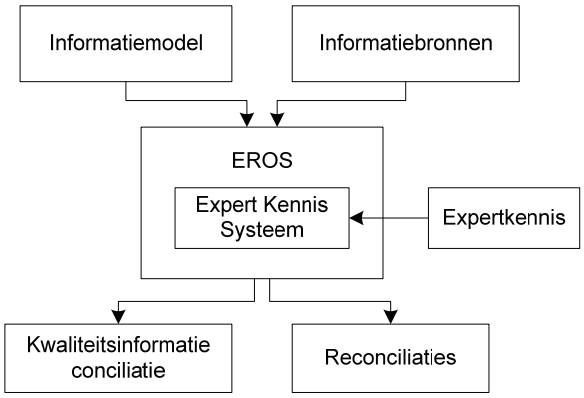
— EROS maakt een keuze in de reconciliatiemogelijkheden door de meest waarschijnlijke mogelijkheden te reconciliëren.
— EROS slaat kwaliteitsinformatie over de conciliatie op.

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

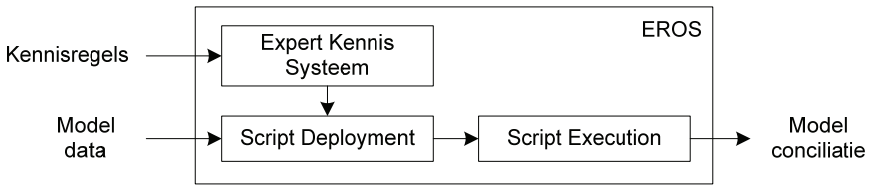
De gebruiker geeft het informatiemodel (inclusief centrumknoop) op. Verder gebruikt EROS de gegevens uit de informatiebronnen om gelijkenis te bepalen. De expertkennis die hierbij nodig is, wordt opgeslagen in een Expert Kennis Systeem. EROS produceert uiteindelijk reconciliaties, maar slaat ook andere reconciliatiemogelijkheden op, alsmede kwaliteitsinformatie. De in- en uitvoer van EROS is schematisch weergegeven in figuur 21.

Figuur 21 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 (Durkin, 1994). Dit valt buiten het onderzoekskader. De systeemarchitectuur van EROS is weergegeven in figuur 22.

Figuur 22 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 gelijkenisdistributie 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.

Eis. 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.

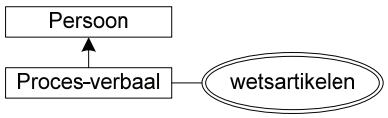
— In de detailview bestaat de primaire sleutel uit de attributen PK_ET en PK_EA.

De detailview kan verder nog een attribuut bevatten dat het aantal voorkomens van een attribuutwaarde bevat. Dit gegeven wordt gebruikt in de reconciliatie (zie paragraaf 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, zie bijlage 1).

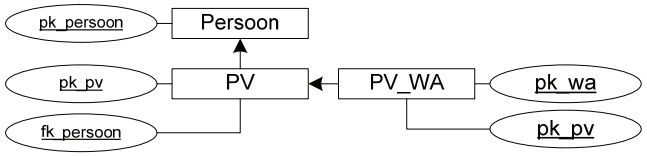
Het afleiden van het informatiemodel uit het fysieke model wordt gedemonstreerd aan de hand van een voorbeeld. Gegeven een hiërarchisch model met personen en processen-verbaal. Een proces-verbaal heeft één of meer wetsartikelen. Het gemeenschappelijke informatiemodel, waarin ook het conceptuele attribuut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

Figuur 23 Wetsartikel als meerwaardig attribuut (conceptueel)



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

Figuur 24 Wetsartikel als meerwaardig attribuut (fysiek)

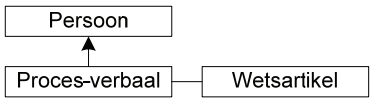


Uit het fysieke datamodel kan het gemeenschappelijke informatiemodel worden afgeleid:
— Een hoofdview, horend bij een entiteittype, heeft één primair attribuut.
— De naam van het bijbehorende entiteittype kan afgeleid worden uit het primaire attribuut of de viewnaam.
— Een detailview heeft twee primaire 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 25 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 24).

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:

Tabel 8 Eigenschappen van een kennisregel

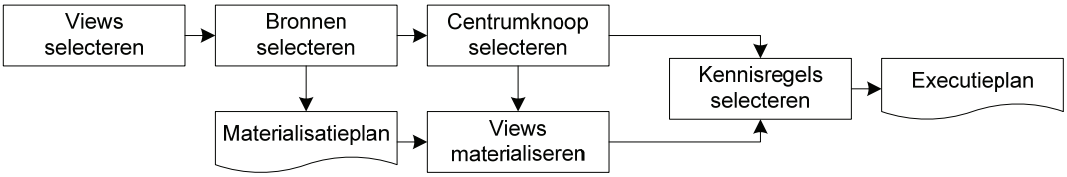
Eigenschap	Omschrijving
naam	De naam van de kennisregel
$bron_1$, $bron_2$	De bronnen waarvoor de kennisregel geldt
knoop	De knoop waar de kennisregel bij hoort
$attr_1$, $attr_2$	De attributen die de gemeenschappelijke eigenschap beschrijven
$aantal_1$, $aantal_2$	De attributen die het aantal voorkomens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

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

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 26 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 gelijkenisdistributie 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:

Tabel 9 Tabelsoorten

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

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 gelijkenismatrices:

Tabel 10 Systeemtabellen

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

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 11.

Tabel 11 Opbouw statistiektabel

Attribuut	Betekenis
PK_I	Verwijzing naar de entiteit in bron I
PK_{II}	Verwijzing naar de entiteit in bron II
Similarity	De gelijkenis 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

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 gelijkenis wordt opgeslagen die kennisregels opleveren. Elke knoop heeft een statistiektabel voor kennisregels.

Tabel 12 Opbouw 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 gelijkenis van de kennisregel

5.4 Implementatie

De implementatie wordt besproken in bijlage E (zie bijlage 1). 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:

Tabel 13 Behaalde resultaten

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

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 (meer hierover in bijlage E, zie bijlage 1), de missende geboortedatums in HKS zijn ingevuld met geboortejaren en de ruisdrempel in de kennisregel “Antecedentdatum vs. Pleegdatum” wordt alleen ingezet als de positie maximaal een jaar is. In de derde en laatste ronde is de bepaling van gelijkenis tussen landentiteiten verbeterd (meer hierover in bijlage D, zie bijlage 1). De verschillende ronden worden uitgebreid besproken in bijlage F (zie bijlage 1).

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.

Tabel 14 Contingentietabel voor resultaten

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

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:

Tabel 15 Indeling incorrecte resultaten

reden fout	fout-positief	fout-negatief
niet aanwezig	onterecht gekozen	n.v.t.[4]
niet gekozen	verkeerd gekozen	niet gekozen
niet gevonden	niet gevonden (fout-positief)	niet gevonden (fout-negatief)

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

[4] 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 gelijkeniswaarde groter dan de gelijkenisdrempel.
Definitie. Een reconciliatiealternatief, of kortweg alternatief, is een andere mogelijkheid dan de correcte mogelijkheid.

Na drie rondes zijn uiteindelijk de volgende resultaten behaald:

Tabel 16 Contingentietabel van de behaalde resultaten

		HKS Gelijk in werkelijkheid? ja		HKS Gelijk in werkelijkheid? nee		OMDATA Gelijk in werkelijkheid? ja		OMDATA Gelijk in werkelijkheid? nee	
Gelijk in EROS?	ja	8.108	81,1%	133	1,3%	8.108	93,1%	133	1,5%
	nee	490	4,9%	1269	12,7%	464	5,3%	0	0,0%

Ingedeeld in goede en foute resultaten geeft dit:

Tabel 17 Goede en foute resultaten

Categorie	HKS		OMDATA	
Goed-positief	8.108	81,1%	8.108	93,1%
Goed-negatief	1.269	12,7%	0	0,0%

Tabel 18 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%

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:

Tabel 19 Aantal reconciliatiemogelijkheden en bijdrage aan het foutpercentage

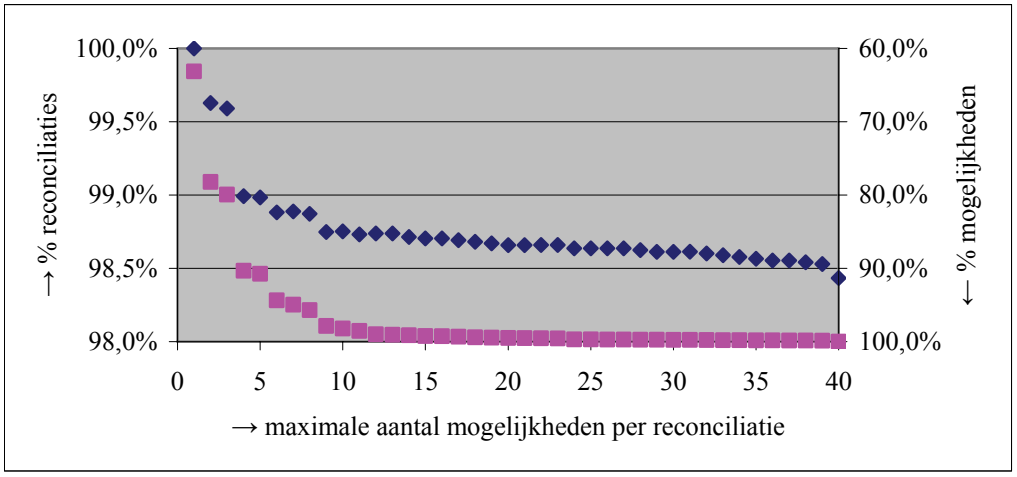
Aantal reconciliatiemogelijkheden	Volume	Bijdrage aan foutpercentage
0	5,3%	5,3%
> 3	20,7%	1,2%
2 of 3	14,3%	0,4%
1	59,7%	0,0%

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 27 Reconciliaties t.o.v. 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.

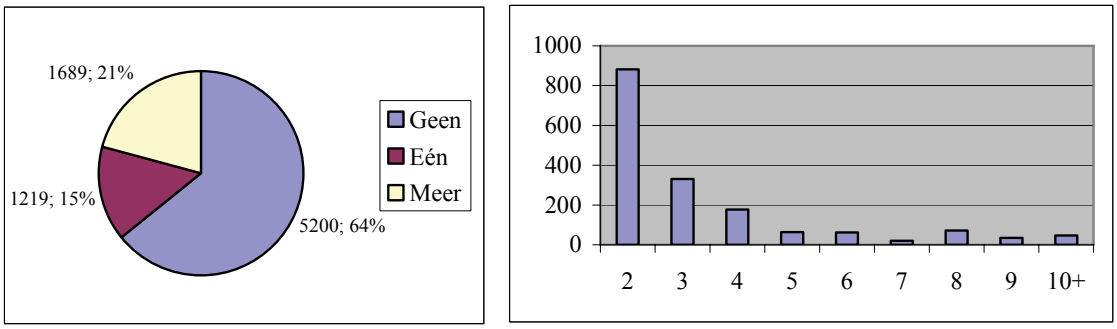
6.3 Kwaliteit

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

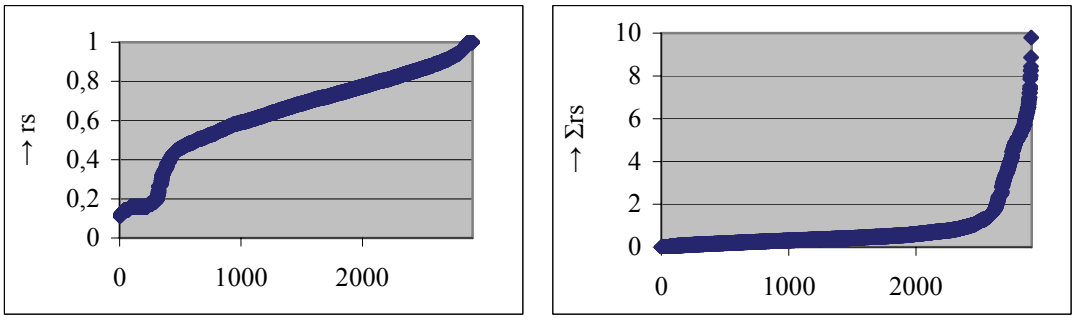
Goed-positief
De waarschijnlijkheid dat een correcte reconciliatie gemaakt wordt, hangt af van het aantal alternatieven dat in het proces gevonden wordt. De grafiek in figuur 28 (links) geeft aan dat bij het merendeel van de reconciliaties (64%) geen alternatief gevonden wordt. Deze reconciliaties hebben de hoogst mogelijke waarschijnlijkheid om gekozen te worden.

Figuur 28 Aantal alternatieven — 3 categorieën (links, absoluut en relatief) en Categorie Meer (rechts, 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 29, links). Maar ook de som van scores Σrs speelt een rol, omdat hiermee de waarschijnlijkheden vanuit elke bron ($prob_{HKS}$ en $prob_{OMDATA}$) van de correcte reconciliatiemogelijkheid afnemen (figuur 29, rechts).

Figuur 29 Grootste dreiging onder de alternatieven (links, *rs*, 1=maximale dreiging) en de som van de dreiging van alle alternatieven (rechts, Σrs, geen maximum)

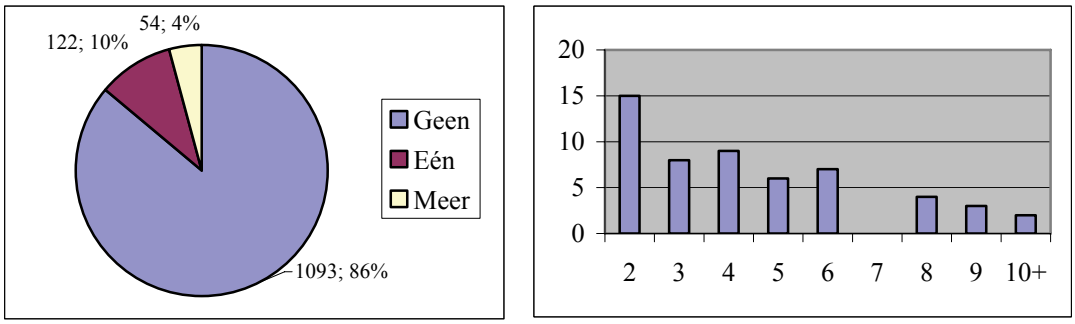


Figuur 29 (links) laat zien dat de dreiging van gevonden alternatieven sterk verschilt per correcte mogelijkheid. Hierover kunnen geen algemene uitspraken gedaan worden, behalve dat de dreiging kan worden verlaagd door te proberen de gelijkenis nog beter te beschrijven. Figuur 29 (rechts) laat zien dat de dreiging van alle alternatieven laag is. Een figuur in deze vorm is wenselijk.

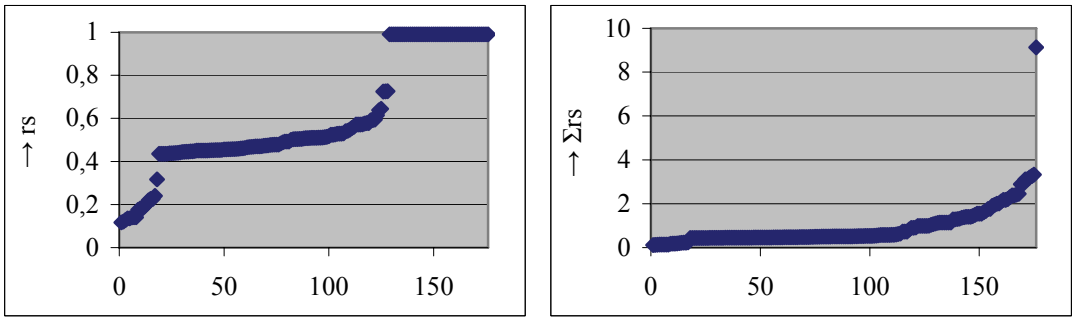
Goed-negatief
Voor deze categorie gelden dezelfde kwaliteitanalyses als voor de categorie ‘goed-positief’. Hier wordt alleen niet gesproken over dreiging van alternatieven,

maar over dreiging van mogelijkheden, omdat in deze categorie geen correcte mogelijkheid bestaat.

Figuur 30 Aantal mogelijkheden — 3 categorieën (links, absoluut en relatief) en categorie Meer (rechts, absoluut)



Figuur 31 Grootste dreiging onder de mogelijkheden (links, *rs*, 1 = maximale dreiging) en de som van de dreiging van alle mogelijkheden (rechts, Σ*rs*, geen maximum)



De dreiging van mogelijkheden is beperkt, zo blijkt uit figuur 30. Een relatief groot aantal gevallen lijkt te maken te hebben met een mogelijkheid met maximale dreiging (figuur 31, links). 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 gelijkenis. 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.

Tabel 20 Subcategorieën in categorie 'niet gevonden'

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

Tabel 21 Afwijkingen in persoonskenmerken

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]

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 22 laat zien.

Tabel 22 Afwijkingen in persoonskenmerken in de 2e ronde

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]

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)'.

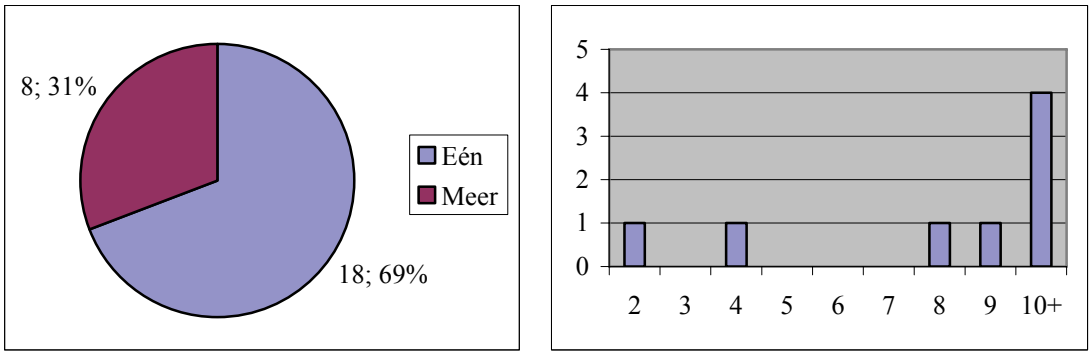
Tabel 23 Resultaat van dreiging in de categorie 'niet gevonden'

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

Tabel 23 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 reconciliatiemogelijkheid wordt gevonden. Daarna kan opnieuw gekeken worden of de correcte reconciliatiemogelijkheid gekozen is.

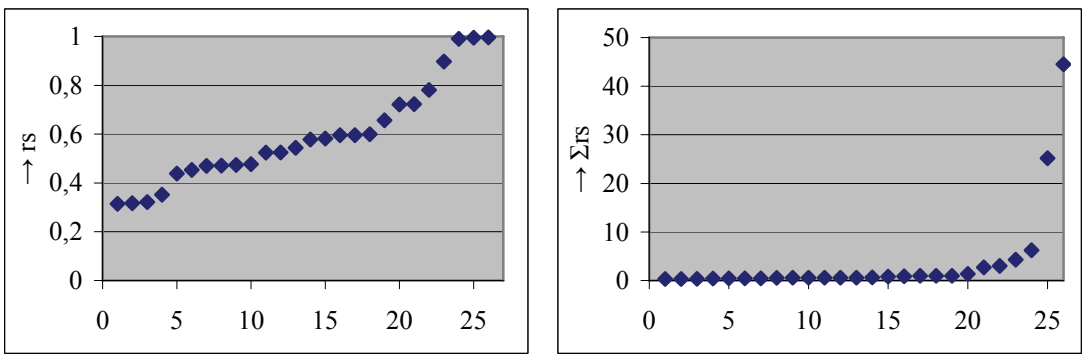
Onterecht gekozen
Bij persoonsentiteiten die gereconcilieerd worden terwijl er geen correcte reconciliatiemogelijkheid is, spelen dezelfde factoren mee als in de categorie 'goed-negatief', met het verschil dat hier een reconciliatiemogelijkheid gekozen is. Daarom zijn ook dezelfde grafieken van toepassing.

Figuur 32 Aantal mogelijkheden —2 categorieën (links, absoluut en relatief) en categorie Meer (rechts, absoluut)



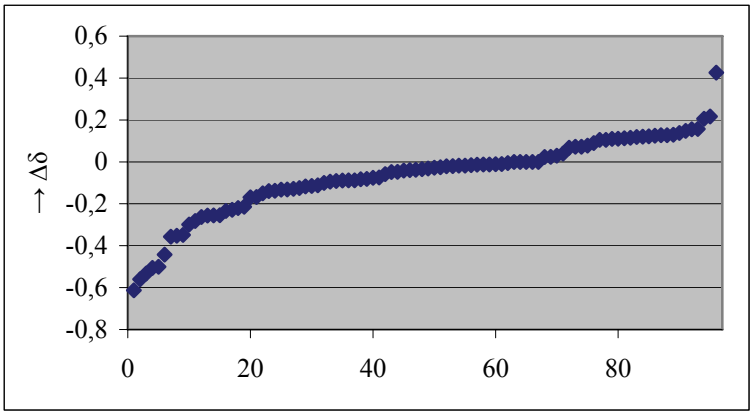
De dreiging bestaat uit de reconciliatiescore (rs) van de gekozen reconciliatie (figuur 33, links) en de som van de score (Σrs) van alle reconciliatiemogelijkheden (figuur 33, rechts).

Figuur 33 Dreiging van de gekozen mogelijkheid (links, *rs*, 1 = maximale dreiging) en de som van de dreiging van alle mogelijkheden (rechts, Σrs, geen maximum)



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

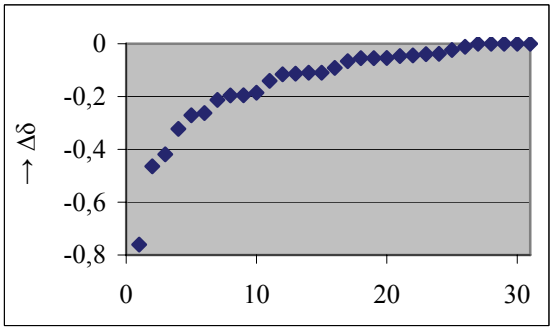
Figuur 34 Positie tussen correcte en gekozen mogelijkheid (correct minus gekozen; [-1,1])



Als de positie 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 positie 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 positie 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 35 Positie tussen correcte en gekozen mogelijkheid (correct minus gekozen; [-1,1])



Net als in categorie ‘verkeerd gekozen’ kan de positie bepaald worden tussen de correcte en de gekozen mogelijkheid. Het blijkt dat voor alle mogelijkheden geldt dat de positie 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 gelijkenis 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?

De overlap tussen twee informatiebronnen wordt gedefinieerd door bronexperts gemeenschappelijke eigenschappen op attribuutniveau te laten beschrijven. Door eigenschappen te beschrijven door middel van positie 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 gelijkenis het beste beschreven. De gelijkenis in elke knoop wordt gedistribueerd naar het centrale gemeenschappelijke entiteittype, waarvoor conciliatie gewenst is: de centrumknoop. De relaties tussen knopen in de gelijkenisdistributie worden beschreven in een informatiemodel. De gelijkenis in een knoop kan efficiënt berekend worden door clustering in een hiërarchisch informatiemodel. De gelijkenis in andere modellen wordt berekend met een gelijkenisstelsel van meerdere informatiemodellen.

De berekende gelijkenis komt samen in de centrumknoop en vormt zo de objectgelijkenis 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% 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 objectgelijkenis. Door de gelijkenisdistributie

kunnen alle gemeenschappelijke eigenschappen die tussen twee informatie-bronnen bestaan effectief gebruikt worden om te bepalen of twee gelijksoortige entiteiten objectgelijk zijn. Mede door gebruik van clustering in de berekening van gelijkenis worden reconciliaties efficiënt berekend, ondanks de beperkte overlap van slechts vijf gemeenschappelijke eigenschappen.

Summary

Entity Reconciliation using Object Similarity

Case 'Matching person entities without the existence a common identifier'

For research purposes a unified and integral view on entities in the field of police and justice is of crucial importance. For several reasons, e.g. enforcement of privacy, linking databases on primary and foreign keys is not always possible or desired. We have developed an approach that focuses on reconciliation in these situations. Our approach is based on exploiting a set of overlapping and related attributes. An attribute in this set does not uniquely identify an entity, but discriminates entities to a certain extent (i.e., the selectivity factor is not too large).

To perform reconciliations, we have combined schema information and the content of databases with available domain knowledge of experts. Schema information of different databases is used to determine what parts of a schema pertain to the same real-world entity. The content of the databases and available domain knowledge are used to define similarity functions. These functions are used to decide whether tuples in different databases refer to the same real-world entity or not.

We have implemented our approach, resulting in a prototype called EROS. We have applied EROS on two databases in the field of police and justice. It appears that our approach can be marked as quite effective, since more than 93% of the tuples have been correctly reconciliated.

For the time-being EROS is able to process only two databases at a time. Extending EROS as such that it is able to process more than two databases at a time in an efficient way is a topic for further research. Extending the knowledge system of EROS with more rules is another topic for further research. To what extent our approach can be generalized is also a topic that needs attention.

Literatuur

Agarwal, S., Keller, A.M., Wiederhold, G., & Saraswat, K. (1995). Flexible Relation: An approach for integrating data from multiple, possibly inconsistent databases. *Proceedings of the 11th International Conference on Data Engineering (ICDE)*, 495-504.

Bonatti, P., Deng, Y. & Subrahmanian, V.S. (2003). An Ontology-Extended Relational Algebra. *Proceedings of the IEEE Information Reuse and Integration (IRI) 2003*, 192-199.

Breitbart, Y., Olson, P.L., & Thompson, G.R. (1986). Database integration in a distributed heterogeneous database system. *Proceedings of the 2nd International Conference on Data Engineering (ICDE)*, 301-310.

Chen, A.L.P., Tsai, P.S.M., & Koh, J.-L. (1996). Identifying Object Isomerism in Multidatabase Systems. *Distributed and Parallel Databases, 4*(2), 143-168.

Choenni, R.S., Blok, H.E., & Fokkinga, M.M. (2004). Extending the Relational Model with Uncertainty and Ignorance. *Technical report TR-CTIT-04-29, juli 2004*, 1-16. Nederland: Centre for Telematics and Information Technology (CTIT), Universiteit Twente.

Choenni, R.S., Blok, H.E., & Leertouwer, E.C. (2006). Handling Uncertainty and Ignorance in Databases: A Rule to Combine Dependent Data. *Proceedings of the 11th International Conference on Database Systems for Advanced Applications (DASFAA),* 310-324.

Cohen, W.W. (1998). Integration of Heterogeneous Databases without Common Domains using Queries Based on Textual Similarity. *Proceedings of the ACM SIGMOD International Conference on Management of Data 1998*, 201-212.

Czejdo, B., Embley, D.W., & Rusinkiewicz, M. (1987). An approach to schema integration and query formulation in federated database systems. *Proceedings of the 3rd International Conference on Data Engineering (ICDE)*, 477-484.

DeMichiel, L.G. (1989). Resolving database incompatibility: An approach to performing relational operations over mismatched domains. *IEEE Transactions on Knowledge and Data Engineering (TKDE), 1*(4), 485-493.

Dey, D., Sarkar, S., & De, P. (1998). A Probablistic Decision Model for Entity Matching in Heterogeneous Databases. *Management Science, 44*(10), 1379-1395.

Dey, D., Sarkar, S., & De, P. (2002). A Distance-Based Approach to Entity Reconciliation in Heterogeneous Databases. *IEEE Transactions on Knowledge and Data Engineering (TKDE), 14*(3), 567-582.

Durkin, J. (1994). *Expert Systems - Design and Development*. Englewood Cliffs, NJ: Prentice Hall International.

Eggen, A.Th.J. & Heide, W. van der (2004). *Criminaliteit en rechtshandhaving 2004*. Den Haag: Boom Juridische uitgevers Onderzoek en beleid 237.

Gass, S.I. (1986). A Process to Determine Priorities and Weights for Large-Scale Linear Goal Programming. *Journal of the Operational Research Society, 37*, 779-785.

Hernández, M.A. & Stolfo, S.J. (1995). The Merge/Purge Problem for Large Databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data 1995*, 127-138.

Hussain, T., Shamail, S., & Awais, M.M. (2003). Eliminating process of normalization in relational database design. *Proceedings of the 7th International Multi Topic Conference (INMIC)*, 408-413.

Kim, W., Choi, I., Gala, S., & Scheevel, M. (1993). On resolving semantic heterogeneity in multidatabase systems. *Distributed and Parallel Database, 1*(3), 251-279.

Lim, E., Srivastava, J., Prabhakar, S., & Richardson, J. (1993). Entity identification in database integration. *Proceedings of the 9th International Conference on Data Engineering (ICDE)*, 294-301.

Lim, E., Srivastava, J., & Shekhar, S. (1996). An evidential reasoning approach to attribute value conflict resolution in database integration. *IEEE Transactions on Knowledge and Data Engineering (TKDE), 8*(5), 707-723.

Monge, A.E. & Elkan, C.P. (1996). The Field Matching Problem: Algorithm and Applications. *Proceedings of the 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 267-270.

Prabhakar, S., Richardson, J., Srivastava, J., & Lim, E. (1993). Instance-Level Integration in Federated Autonomous Databases. *Proceeding of the 26th Hawaii International Conference on System Sciences (HICSS)*, vol. iii, 62-69.

Scott, A.C., Clayton, J.E., & Gibson, E.L. (1991). *A Practical Guide to Knowledge Acquisition*. Reading, MA: Addison-Wesley.

Silberschatz, A., Korth, H.F., & Sudarshan, S. (2005). *Database System Concepts, 5th Edition*. Boston, MA: McGraw-Hill.

Stefik, M. (1995). *Introduction to Knowledge Systems*, San Fransisco, CA-: Morgan Kaufmann.

Templeton, M., Brill, D., Dao, S.K., Lund, E., Ward, P., Chen, A.L.P., & MacGregor, R. (1987). Mermaid - A front end to distributed heterogeneous databases. *Proceedings of the IEEE, 75*(5), 695-708.

Wang, Y.R. & Madnick, S. (1989). The interdatabase instance identification problem in integrating autonomous systems. *Proceedings of the 5th International Conference on Data Engineering (ICDE)*, 46-55.

Bijlage 1 Overzicht bijlagen

In dit memorandum wordt verwezen naar verschillende bijlagen. Deze bijlagen zijn opvraagbaar bij de auteur.

Bijlage A	Informatiebronnen
Bijlage B	Literatuuronderzoek
Bijlage C	Theoretische achtergrond
Bijlage D	Voorbereiding casus
Bijlage E	Implementatie
Bijlage F	Presentatie resultaten
Bijlage G	Extrapolatie

Naast deze bijlagen horen bij dit memorandum de volgende documentbijlagen. De documentbijlagen 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 het memorandum (architectuur, ontwerp en implementatie) bevat het document ook het testplan en documentatie voor gebruik van het prototype.