



Universiteit
Leiden
The Netherlands

Aspects of Record Linkage

Schraagen, M.P.

Citation

Schraagen, M. P. (2014, November 11). *Aspects of Record Linkage*. Retrieved from <https://hdl.handle.net/1887/29716>

Version: Corrected Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/29716>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/29716> holds various files of this Leiden University dissertation.

Author: Schraagen, Marijn Paul

Title: Aspects of record linkage

Issue Date: 2014-11-11

Appendix C

Samenvatting in het Nederlands

Dit proefschrift behandelt verschillende aspecten van het *record linkage-probleem*. Een record (in goed Nederlands *dossier* geheten) is een vermelding in een database, bijvoorbeeld een gebruikersprofiel op een webpagina of een geboortevermelding in de burgerlijke stand. Binnen een database komt het regelmatig voor dat meerdere records voor dezelfde entiteit aanwezig zijn, bijvoorbeeld twee gebruikersprofielen voor dezelfde persoon met een verschillend e-mailadres. Ook bevatten de meeste databases records over verschillende onderwerpen (personen, gebeurtenissen, objecten etc.) die gerelateerd aan elkaar zijn, zoals twee geboortes met dezelfde ouders. Het is in veel gevallen niet triviaal om te bepalen welke records kunnen worden gekoppeld en welke niet, dit is het probleem van record linkage (dat overigens gegeneraliseerd kan worden naar meerdere databases).

Het proefschrift bevat verschillende deelonderzoeken uitgevoerd met de *Genlias-database*. Dit is een historische database met akten van de burgerlijke stand in Nederland uit (voornamelijk) de 19^e eeuw. Deze data, in totaal ongeveer 15 miljoen aktes, is gedigitaliseerd en beschikbaar gesteld voor onderzoek¹. De digitalisatie is echter nog niet compleet, de database zoals gebruikt in het onderzoek bevat naar schatting ongeveer de helft van alle aktes die in de archieven aanwezig zijn. Dit bemoeilijkt het zoeken

¹De database is vanaf 2002 vrij te doorzoeken op internet. Sinds medio 2012 is de database onder de naam *WieWasWie* beschikbaar op de website <http://www.wiewaswie.nl>.

naar gerelateerde of dubbele records. Daarnaast is de evaluatie van links (koppelingen) een probleem, omdat er geen informatie beschikbaar is over de juistheid van links in de Genlias-database of een deel daarvan. Ook voor vergelijkbare databases zijn er vrijwel geen geverifieerde links beschikbaar.

Vanwege de incomplete digitalisatie en het ontbreken van geverifieerde links wordt er in het proefschrift aandacht besteed aan het voorspellen van de verwachte hoeveelheid links voor een record (Hoofdstuk 3). De methode maakt gebruik van automatisch bepaalde eenvoudige links tussen records om een voorspelling te doen over de beschikbaarheid van aanvullende links voor een record. De netwerkstructuur (in de informatica meestal een *graaf* genoemd) die ontstaat door de eenvoudige links tussen records blijkt voldoende informatie te bevatten om nieuwe links te kunnen voorspellen, zonder hierbij gebruik te hoeven maken van de specifieke inhoud van de records zelf.

Een veelgebruikte methode voor het vinden van links is het vergelijken van records op syntactisch niveau, meer specifiek het verschil in spelling tussen de tekst van de records uitgedrukt in het aantal letters dat moet worden veranderd om de spelling van twee records aan elkaar gelijk te maken. Voor de paarsgewijze vergelijking van records zijn efficiënte methodes beschikbaar, waarmee links kunnen worden gevonden door eenvoudig alle mogelijke combinaties van records met elkaar te vergelijken. Echter, het totale aantal combinaties van records is het kwadraat van het aantal records, waardoor het voor grote databases praktisch onuitvoerbaar is om alle combinaties te controleren. In Hoofdstuk 4 wordt een methode beschreven om alle relevante combinaties te vinden zonder dat alle mogelijke combinaties moeten worden geprobeerd, door het gebruik van een index op de inhoud van de records. In deze index worden records in een bepaalde volgorde geplaatst waarbij alle records die een klein verschil in spelling hebben gegarandeerd bij elkaar in de buurt staan. De efficiëntie van deze index wat betreft zowel rekentijd als geheugengebruik maakt het gebruik van record linkage in real-time toepassingen mogelijk, naast het in relatief korte tijd verwerken van volledige datasets.

De belangrijkste bron van informatie voor het linken van records wordt gevormd door persoonsnamen. Deze zijn vaak erg specifiek en daardoor identificerend voor een bepaald persoon, zeker in combinatie met andere persoonsnamen in een record zoals de namen van de ouders. Echter, vanwege de specificiteit is ook de hoeveelheid variatie in namen erg groot. Het onderzoek gepresenteerd in Hoofdstuk 5 richt zich op een model

dat de variatie in persoonsnamen beschrijft. Dit model leidt op basis van voorbeelden van naamvariatie automatisch een beslisstrategie af waarmee kan worden bepaald welke letters in een naam tot de *kern* van deze naam horen en welke letters deel uitmaken van de variatie tussen namen. In Hoofdstuk 6 wordt naamvariatie opnieuw bekeken vanuit een kwantitatieve benadering. Omdat de Genlias-database relatief veel records bevat, kan een grote hoeveelheid naamvarianten worden verzameld door te kijken welke varianten voorkomen in eenvoudige links. Deze varianten kunnen als zodanig worden gebruikt bij het zoeken naar aanvullende links, of de varianten kunnen (al dan niet automatisch) worden afgebeeld op een beperkte hoeveelheid basisnamen waarbij twee records als link kunnen worden beschouwd als de basisnamen met elkaar overeenkomen. In het onderzoek is echter gebleken dat deze methode in een klein maar significant aantal gevallen onjuiste naamvarianten oplevert, waardoor verschillende controlemechanismen moeten worden gebruikt om tot een bruikbaar resultaat te komen.

In record linkage is het gebruikelijk dat twee records met elkaar vergeleken worden, waarna een link wordt vastgesteld of verworpen aan de hand van de uitkomst van de vergelijking. Echter, in veel gevallen kunnen andere records die niet direct betrokken zijn bij de vergelijking informatie bevatten over de correctheid van een link. Een voorbeeld is een database met wetenschappelijke artikelen, waarbij het vaak onzeker is of twee artikelen waarbij de naam van de auteur gelijk is ook door dezelfde persoon geschreven zijn. Echter, als er bij beide artikelen een (van elkaar verschillende) co-auteur betrokken is en er kan een derde artikel worden gevonden waar deze co-auteurs gezamenlijk worden vermeld, dan maakt dit derde artikel een link tussen de twee originele artikelen veel waarschijnlijker. Ook in de burgerlijke stand kunnen dergelijke situaties optreden, bijvoorbeeld een link tussen twee geboorteaktes die als incorrect kan worden beoordeeld door de aanwezigheid van een overlijdensakte van de moeder in de periode tussen de twee geboortes. Hoofdstuk 7 beschrijft een onderzoek naar het gebruik van dergelijke informatie.

In Hoofdstuk 7 wordt een benchmark-database gebruikt, dit is een externe database met (gedeeltelijk) dezelfde records als in Genlias waarbij links tussen records handmatig zijn toegevoegd. Deze database heeft een ander formaat dan Genlias, waardoor er eerst moet worden bepaald welke records met elkaar overeenkomen voordat de links kunnen worden vergeleken. Het gebruik van verschillende databases wordt verder behandeld

in Hoofdstuk 8. Het onderzoek richt zich daarbij op het onder genealogen populaire Gedcom-bestandsformaat, wat in opzet sterk afwijkt van de Genlias-database. De verschillen in structuur bemoeilijken de vergelijking tussen records, maar deze aanpak heeft als voordeel, naast het vergroten van het toepassingsgebied van de verschillende algoritmes, dat er in de vorm van bestaande genealogische bestanden een grote hoeveelheid extra informatie kan worden gebruikt voor het vinden en verifiëren van links.

In het proefschrift wordt onderzocht hoe de computer kan worden gebruikt bij het koppelen van grote hoeveelheden records. Hierbij is gebleken dat deze taak niet eenvoudig te automatiseren is. Echter, mensen zijn over het algemeen goed in staat om te beoordelen of twee records aan elkaar gerelateerd zijn, en bij het zoeken naar een link is de strategie van mensen om snel tot een resultaat te komen vaak succesvol. Als onderdeel van het onderzoek is daarom ook vanuit een cognitief psychologisch perspectief gekeken naar het gedrag van mensen bij het uitvoeren van een taak die kan worden gebruikt tijdens het uitvoeren van record linkage. Deze taak (beschreven in Hoofdstuk 9) betreft het onderverdelen van persoonsnamen in een stam en een suffix (in de brede betekenis van het woord *suffix*, d.w.z. een of meerdere letters aan het eind van een woord), bijvoorbeeld de naam *Beekman* verdeeld in *Beek* en *man*. De vraag in het onderzoek was of veelgebruikte suffixes (zoals *man*) sneller worden herkend dan weinig gebruikte suffixes (bijvoorbeeld *bor*). Uit het onderzoek is niet gebleken dat er in het algemeen een verschil is tussen de twee suffixtypes. Dit kan verschillende oorzaken hebben, zoals interferentie van andere processen tijdens het onderverdelen van een naam. Om meer duidelijkheid te krijgen over de precieze cognitieve processen is verder onderzoek noodzakelijk.

De deelonderzoeken in dit proefschrift behandelen verschillende aspecten van het record linkage-probleem. Dit is uiteraard slechts een selectie van mogelijke aspecten en uitgangspunten. Echter, als deze selectie in staat blijkt om enige kennis en inzichten te verschaffen op het gebied van record linkage, dan is het proefschrift in zijn doel geslaagd.