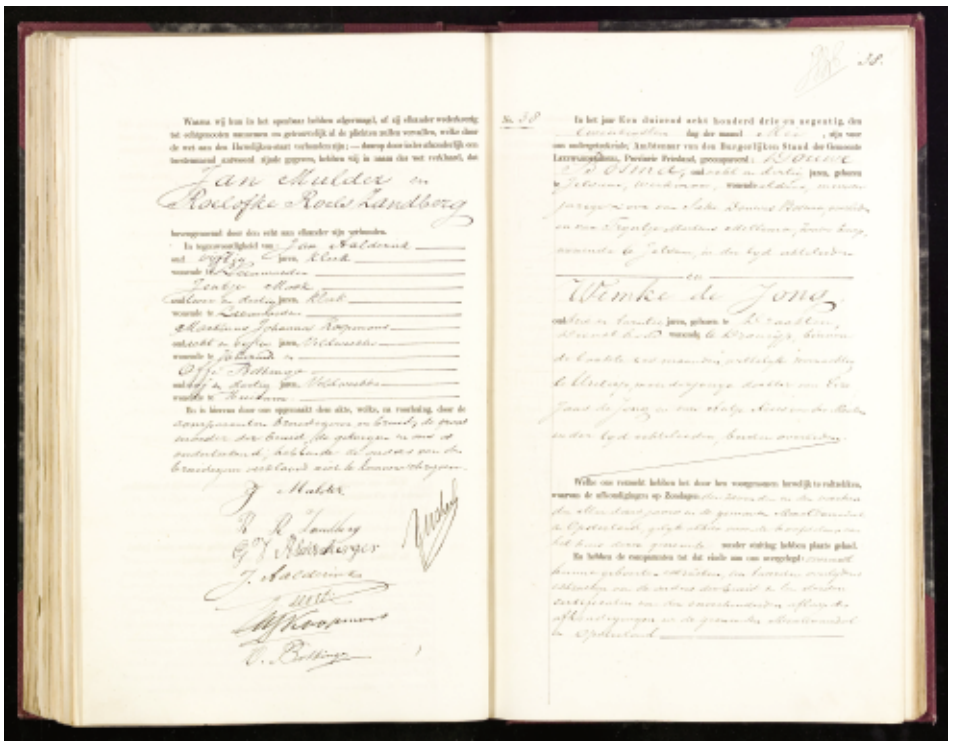
Eén basisregistratie voor historische personen

Miljoenen geboorte-, overlijdens- en huwelijksaktes gelinkt

Vanaf 1812 houdt men bij wie wordt geboren, huwt en overlijdt. Deze akten zijn nu gelinkt beschikbaar voor de wetenschap en amateurgenealogen via wiewaswie.nl.

Mathilde Jansen

De aktes van de Burgerlijke Stand zijn een schat aan data voor iedereen die meer wil weten over zijn voorouders. Dat de aktes ook voor demografisch en ander historisch onderzoek relevant zijn, bedacht Kees Mandemakers (IISG) al enkele jaren geleden. Hij bracht een samenwerking tot stand tussen het Centraal Bureau voor Genealogie (CBG) en het IISG. Het CBG leverde de aktes – die vanaf de jaren negentig waren gedigitaliseerd en getranscribeerd door vele vrijwilligers – en onderzoekers van onder andere het IISG ontwikkelden een systeem om de aktes automatisch te linken.



Online is deze huwelijksakte van Douwe Bosma, registratiedatum 1893, nu ook gelinkt aan andere informatie, zoals informatie over de vader en moeder van de bruid en van de bruidegom. *Credits: wiewaswie.nl/nl/detail/7174744*

Minder tijdrovend

Onlangs werd via CLARIAH PLUS een nieuwe methode ontwikkeld waardoor het berekenen van een linkset veel minder tijdrovend werd. Auke Rijpma, projectleider bij CLARIAH PLUS en onderzoeker aan de Universiteit Utrecht en het IISG: “De regelsets die door het team van Mandemakers zijn ontwikkeld, gebruiken we nog steeds. Joe Raad van de Vrije Universiteit heeft een nieuwe implementatie gedaan, gericht op efficiëntie. Nu kost het linken ongeveer een dag.”

De methode werkt op basis van regelsets, oftewel *rule based linking*. De namen in twee verschillende aktes mogen bijvoorbeeld aan elkaar gelinkt worden wanneer ze maximaal twee karakters van elkaar verschillen, mits de naam vijf letters of meer bevat. Namen hoeven niet identiek te zijn, dit heeft bijvoorbeeld te maken met de variatie in schrijfwijze.

Overlinking

Door de grote hoeveelheid aktes gaat het linken nog niet altijd goed. Om de mismatches eruit te halen, kijken de onderzoekers naar ‘overlinking’: “Een geboorteakte moet uniek gelinkt zijn aan de aktes van de ouders. Als een akte vaker gelinkt wordt, gaat er ergens iets mis. Daarnaast maken we een evaluatiedataset waarin we een paar duizend links handmatig aanleggen. Vervolgens maken we een vergelijk met de geautomatiseerde links.”

44 miljoen aktes

Voorlopig zijn alle 12 miljoen huwelijksaktes gelinkt, daar komen nog 20 miljoen geboorteaktes en 12 miljoen overlijdensaktes bij. De gelinkte data zijn om diverse redenen belangrijk: het biedt onderzoekers bijvoorbeeld de mogelijkheid om sociale ongelijkheid te zoeken, zowel in termen van waar je in de maatschappij terechtkomt, als hoe je sociale omgeving eruitziet. “Het is een kwestie van dezelfde procedure opnieuw implementeren. We verwachten dan ook dat dit nog sneller gaat. Ons streven is om nog dit jaar een besloten release voor geïnteresseerde onderzoekers te hebben, en medio 2021 een volwaardige open release.” De gelinkte data zijn al zichtbaar op wiewaswie.nl en komen via druid.datalegend.net beschikbaar voor onderzoekers.

druid.datalegend.net

Schatting van tekstverlies met methodes uit ecodiversiteit

Het Atlantis van de ridderepiek

Letterkundigen duiden literatuur uit de middeleeuwen aan als ‘de verloren gegane wereld Atlantis’ of als ‘wrakhout van een aangespoelde vloot’. De grote vraag is: hoe groot was die vloot? Folgert Karsdorp en Mike Kestemont rekenden het uit.

Mathilde Jansen

Een eerdere ruwe schatting van historisch letterkundige Frits van Oostrom telde ruim honderd verschillende ridderepische teksten. Hij baseerde dit aantal op bestaande bronnen die verwijzen naar teksten die we niet kennen, de tekstgetuigen. Folgert Karsdorp (Meertens Instituut) en Mike Kestemont (Universiteit Antwerpen) besloten om de vraag te beantwoorden met een *resampling methode*, een methode nog niet eerder toegepast binnen historisch letterkundig onderzoek.

Soortenrijkdom

Karsdorp licht toe: “Biologen bepalen aan de hand van observaties het werkelijke aantal diersoorten. Hierbij speelt een negatieve bias: het aantal getelde diersoorten is een onderschatting van de totale soortenrijkdom. Hetzelfde geldt voor middeleeuwse teksten. Door deze overeenkomst kwamen we op het idee om gebruik te maken van de statistische methoden die biologen gebruiken om de totale soortenrijkdom te bepalen.” De onderzoekers lieten zich in hun methodiek inspireren door Alex Mesoudi, hoogleraar Culturele Evolutie aan de Universiteit van Exeter (UK), die al in 2011 een pleidooi hield voor een Darwiniaanse studie van de ontwikkelingsgang van culturele fenomenen zoals literatuur.

Nieuwe conclusie

Karsdorp en Kestemont gingen bij hun berekening uit van 74 unieke teksten en 164 tekstgetuigen. De eerste toegepaste methode was *Jackknife resampling* (knipmesmethode). Hierbij wordt elke tekstgetuige tijdelijk uit de data verwijderd. Op basis van deze methode werd een schatting gemaakt van (minimaal) 113 tot (maximaal) 128 teksten. De tweede methode was de methode-Chao, die op een verregaande manier gebruikmaakt van *resampling*. De schatting met behulp van deze methode kwam op 106 tot 219 teksten. Karsdorp: “Op basis van de uitkomsten is de schatting van Van Oostrom niet onmogelijk, maar wel erg conservatief. Volgens de laatste methode is het immers ook mogelijk dat we met 219 originele teksten te maken hebben gehad. Dat zou betekenen dat zelfs maar de helft van het totaal aantal teksten bewaard is gebleven.”

Karel ende Elegast is een ridderverhaal met een hoofdrol voor Karel de Grote, geschreven omstreeks 1270. Dankzij meerdere drukken uit de 15de en 16de eeuw is dit de enige Middelnederlandse Karelroman waarvan de volledige tekst beschikbaar is. *Credits: KB*

DOI: 10.2143/SDL.61.3.3287540

IISG-collectie doorzoekbaar in meerdere talen

Het IISG heeft veel boeken, tijdschriften en archiefmateriaal in latijnse talen en alfabetten, zoals Amhaars, Arabish, Bengali, Farsi en Tamil. Om dat materiaal te catalogiseren, is altijd gebruikgemaakt van transliteratie – het ‘converteren’ van de niet-westerse karakters naar het westerse alfabet.

Catalogussystemen ondersteunden vroeger alleen het westerse alfabet. Bij transliteratie kan echter veel fout gaan. Bovendien was het materiaal op deze manier slecht vindbaar voor mensen uit de taalgebieden waar de collecties vandaan komen. Daarom werden tijdelijk vijf *native speakers* aangesteld die auteursnamen, titels en archiefbeschrijvingen aanleverden van tijdschriften en boeken die al in de catalogus zitten. Het resultaat staat op de website van het IISG. (MJ)

iisg.amsterdam/nl/collecties