Anne Frank in Bunschoten-Spakenburg
Pilot Geocoderen Netwerk Oorlogsbronnen

Menno den Engelse en Annelies van Nispen ■

Hoe kun je meer dan acht miljoen oorlogsbronnen automatisch van geografische informatie voorzien? Dat onderzochten we voor Netwerk Oorlogsbronnen (NOB). Door geocodes aan metadata te koppelen zijn de bronnen bijvoorbeeld via een kaart te ontsluiten. Hoewel dat koppelen aardig lukte, is het duidelijk dat zo'n automatisch proces niet probleemloos is. Vooral de veelsoortige metadata speelden ons parten. We concludeerden dat het gebruik van URI's in eigen collectiebeheersystemen veel kan oplossen.

De NOB-portal biedt toegang tot WO2-bronnen uit bijna dertig verschillende collecties. NOB is het best te omschrijven als een thematische aggregator. Sommige collecties zijn in hun geheel opgenomen, zoals de Beeldbank van het NIOD. Andere gedeeltelijk, zoals bronnen van het Nationaal Archief die de oorlogsjaren betreffen. NOB filtert en *harvest* ook landelijke aggregatoren als DiMCoN en Digitale Collectie op het thema Tweede Wereldoorlog. Er worden collecties van zowel archieven, musea als bibliotheken opgenomen, die elk hun eigen beschrijvingsstandaarden en praktijken meebrengen. Logischerwijs zijn de metadata – en de kwaliteit ervan – dus zeer divers.

Geografische metadata

De eerste stap in de pilot betrof een onderzoek naar de kwaliteit van de bestaande geografische metadata van de bronnen uit het NOB-portal. Geografische termen troffen we in vier verschillende Dublin Core-velden aan. Het vrolijkst werden we van termen in een *dc:coverage*-veld. Weliswaar slaat 'dekking' zowel op geografie als op tijd, maar de enkele datum of periodeaanduiding was goed te herkennen. We konden er dus van op aan dat een beschrijving in dit veld een geografische term was.
Een Uniform Resource Identifier (URI) hebben we nergens als term aangetroffen. Helaas, want URI's identificeren niet alleen eenduidig, maar brengen ook coördinaten en andere informatie over locaties binnen handbereik. Een URI blijft door de tijd heen hetzelfde, ongeacht of de naam van een plaats verandert. Zo kan de geografische locatie altijd worden teruggevonden op het internet. Er is geen afhankelijkheid meer van spelling of hiërarchie.

Oorlogsbronnen Maartensdijk op de kaart.

Vaak werd in een of meerdere *coverage*-velden een hiërarchie vermeld, waarbij de volgorde niet altijd dezelfde was. 'Plaats/provincie/land' dus, maar ook 'land/provincie/plaats', 'plaats/land' of 'plaats/gemeente/provincie/land'. Eilanden als Walcheren en Java wisten zich soms ook in de hiërarchie te werken. Duidelijk voor een mens, maar taaie kost voor een geocoder.

Lastig

Hoewel je van het veld *dc:subject* weet dat het veld trefwoorden bevat, is het voor een script lastig te bepalen of het om een *geografisch* trefwoord gaat. Je moet dan van elk trefwoord kijken of een geocoder het als locatie herkent. Dankzij straatnaamcommissies die zichzelf te modern vinden voor ouderwetse achtervoegsels als -straat, -laan of -weg, krijg je dan soms het bevestigende antwoord dat Anne Frank een straat is, zoals in Bunschoten-Spakenburg.
Veel beschrijvingen ontberen de twee bovengenoemde velden. Vooral in archiefcollecties, alleen al door omvang vaak minder uitvoerig gemetadateerd, zijn geografische aanduidingen alleen in de tekstvelden *dc:title* en soms *dc:description* te vinden. Voor

De Dam met Duitse richtingborden (coll. NIOD).

Een overzicht van gevonden vals-positieven. Zie ook: bit.ly/IslandsofMeaning.

de krantenartikelen, met meer dan zeven miljoen de bulk in het NOB, geldt hetzelfde.
In deze gevallen zijn de tekstvelden doorzocht met de combinatie *partition of speech (POS) tagger* / reguliere expressie. De POStagger analyseert de zinsbouw en vindt eigennamen, ook als ze uit meerdere woorden zijn opgebouwd: New York, Den Haag en dus ook Anne Frank. De reguliere expressie kijkt of de eigennaam wordt voorafgegaan door 'in', 'te', 'naar' et cetera.

Geocoderen

In de volgende stap zijn de iets meer dan 25.000 (vermoedelijk) geografische termen uit het NOB-portal tegen geografische thesauri gehouden. Eerst tegen de Historische Geocoder. Waar die geen resultaat gaf vervolgens ook tegen de GeoNames API. De Historische Geocoder beperkt zich grotendeels tot Nederland, maar kent anders dan GeoNames ook straten en adressen en koppelt veel historische plaatsnaamvarianten aan GeoNames URI's.
Ongeveer tienduizend termen wisten we zo eenduidig te geocoderen. Tweeduizend termen gaven meerdere resultaten – naast Zeeland kent bijvoorbeeld ook Zuid-Holland een Middelburg. Zelfs van een uitstekend gemetadateerde set als de NIOD-beeldbank konden we automatisch maar zo'n 65% eenduidig oplossen. Met enig handwerk kregen we dat uiteindelijk tot zo'n 90%. Dertienduizend termen gaven geen enkel resultaat. Niet altijd onterecht overigens: termen als arbeidsinzet, illegaliteit en zondagmiddagcabaret waren ook uit tekstvelden geplukt.
Vanwege het open karakter, de wereldwijde dekking, de gebruiksvriendelijke API van GeoNames zelf en de brede toepassing van GeoNames URI's in wetenschap en het erfgoedveld, hebben we zo veel mogelijk GeoNames URI's opgenomen. Voor straten, adressen en gebouwen zijn BAG-id's (Basisadministratie Adressen en Gebouwen) gebruikt. Met die URI's en id's was het eenvoudig coördinaten en namen in elke gewenste taal en hiërarchie binnen te halen.
Historische namen waren soms lastig: Sovjet-Unie, Joegoslavië, Nederlands-Indië, Kamp Westerbork en Oranjehotel werden niet door GeoNames herkend. Deze gevallen hebben we aan GeoNames toegevoegd, zodat we alsnog een URI hadden en een volgend persoon geen problemen meer hoeft te verwachten.

Resultaten

We hebben binnen de tienduizend eenduidig gegeocodeerde termen een steekproef gedaan, waarbij zowel gekeken is naar uit coverage afkomstige termen als naar termen uit tekstvelden (titel en beschrijving).
Van die tweehonderd uit coverage afkomstige termen was 0,5% verkeerd gegeocodeerd – de op Nederland gerichte Historische Geocoder zag 'Atlantische Oceaan' als straat in Naaldwijk.
We willen het niet steeds over creatieve straatnaamcommissies hebben, maar buiten de steekproef kwamen we in Purmerend een wijkje tegen met straatnamen als Bali, Kalimantan en Sulawesi. Dat in Suriname een Berlijn ligt, kunnen we hun niet verwijten.
Bij tweehonderd uit tekst afkomstige termen was zo'n 18,5% vals-positief. Daarbij maakt het wel uit of je naar termen kijkt die alléén in tekst voorkomen (vijfentwintig van de honderd vals-positief) of termen die in tekst zijn aangetroffen, maar ook in coverage voorkomen (twaalf van de honderd vals-positief). Naast Anne Frank zijn er straten met namen als Zuidfront, Verzet, Bevrijding en Spitfire. En wereldwijd kwamen we plaatsen tegen als Shakespeare, Axis, Brief en Uniform – van de leukste hebben we een False Positives-wereldkaart gemaakt.

Eigen collectie eerst

De pilot heeft aangetoond dat je een eind komt met automatische processen, maar dat je zonder semihandmatige ingrepen ook in gunstige gevallen maar twee derde oplost. Alleen door eenduidige identifiers – liefst URI's – te gebruiken, is dit naar de honderd procent te brengen.
Het is dus zaak die URI's in het eigen collectiebeheersysteem onder te brengen. Mocht u dat willen: de resultaten van deze pilot zijn per collectie te downloaden van Github (zie: bit.ly/GithubNOBGeo). Permalinks lijken nog steeds bij elke leverancierswisseling te wijzigen, waarna de resultaten vaak niet meer te koppelen zijn. En dan ben je nergens meer.

Info

Meer weten? Lees het nieuwsbericht over de afronding van het project via bit.ly/NOBgeocoding. En lees via bit.ly/IslandsofMeaning een blog over de leukste vals-positieven. ■

Menno den Engelse ■ dataprogrammeur Islands of Meaning.
Annelies van Nispen ■ informatiespecialist Collecties NIOD, Instituut voor Oorlogs-, Holocaust- en Genocidestudies.

Achtergrond

Netwerk Oorlogsbronnen wil de verspreide collecties over en uit de Tweede Wereldoorlog beter vindbaar en bruikbaar maken. Daartoe ontwikkelt NOB een aantal digitale basisdiensten op vier terreinen: wie, wat, waar en wanneer. Centraal hierbij staat dat deze diensten de instellingen beter in staat stellen hun publiek te bedienen. De pilot geocoding is gericht op de waar-vraag: 'zoeken op locatie' door grote hoeveelheden data. Kijk op www.oorlogsbronnen.nl voor meer informatie over NOB.