Eindrapportage Pilot Geocoderen Oorlogsbronnen

De Pilot geocoderen Oorlogsbronnen is uitgevoerd door Menno den Engelse in opdracht van Netwerk Oorlogsbronnen. Projectleider was Annelies van Nispen. De looptijd van de pilot was maart-juni 2016.

Doel en opdracht

Het doel van de pilot Geocoderen Oorlogsbronnen was tweeledig:

- Een data-analyse van de huidige geografische metadata in portal oorlogsbronnen. Hoe is de kwaliteit van de metadata en welke (niet)bruikbare resultaten levert geocoderen van de metadata op,

- een goede en praktische wijze vinden om de door Oorlogsbronnen geharveste metadata te geocoderen en te verrijken. Deze wijze moet aansluiten bij de bestaande technische infrastructuur van oorlogsbronnen.

Met de resultaten van de pilot moet ‘zoeken op plaats/geografisch zoeken’ verbeterd worden, het moet mogelijk worden om ‘hiërarchisch’ geografisch te zoeken (plaats/gemeente/provincie). Er moeten sets gemaakt kunnen worden van geografische eenheden en metadata zal ook via de kaart ontsloten kunnen worden..

De pilot beperkt zich tot geografische aanduidingen in de metadata die wordt geleverd door de partners van NOB. In de toekomst zullen mogelijk ook de OCR-data kunnen worden gebruikt.

De pilot zal onderzoeken welke geocoders gebruikt kunnen worden voor geocoderen. Op dit moment worden de Historische Geocoder, Geonames en TGN genoemd. TGN en vooral GeoNames zijn veelgebruikte geografische thesauri, de Historische Geocoder bevat meerdere geografische thesauri - naast het Nederlandse deel van GeoNames en TGN bijvoorbeeld ook de Basisregistraties Adressen en Gebouwen (***BAG***).
In de pilot moet onderzocht worden welke services gebruikt kunnen worden en wat de voor- & nadelen zijn.

Waar en wat opslaan?

Belangrijk is dat de originele data in originele staat blijft en dat de verrijkingen apart worden opgeslagen. Het onderscheid verrijkt en originele data moet duidelijk blijven.

De verrijkingen zullen bestaan uit identifiers (URI's), geometrie en hiërarchie. Gekeken moet worden welke thesauri (GeoNames, BAG, TGN, Gemeentegeschiedenis) gebruikt zullen worden.

Geocoderen inbouwen in infrastructuur

Onderdeel van de pilot is ook te kijken hoe geocoderen in het huidige harvest & publicatie proces van Oorlogsbronnen kan worden ingepast. De technische infrastructuur wordt onderhouden door Trifork. De geografische metadata & georefereren moeten aansluiten bij deze infrastructuur.
Onderzocht is welke stappen in het proces aangepast of toegevoegd moeten worden, dit voor zowel nieuwe als gewijzigde objecten. De context, bijvoorbeeld uit welke collectie het object komt, is belangrijk en moet meegenomen worden.

NOB op de Kaart

Om nut en noodzaak van het geocoderen te illustreren worden één of meerdere kaartapplicaties gemaakt. Deze kaarten kunnen ook goed gebruikt worden bij het beoordelen en duiden van de resultaten.Het presenteren van bronnen op kaart is een belangrijke uitkomst.

.

De data

De data bestaat uit door Netwerk Oorlogsbronnen geaggregeerde metadata van 27 collecties. Dit kunnen zowel museale, archief als bibliotheekcollecties zijn. Twee collecties (DIMCON en DIGCOL) bestaan op hun beurt weer uit verschillende collecties.

De kwaliteit van de metadata loopt sterk uiteen en dat kan mogelijk de kwaliteit van het geocoderen en geografische toepassingen beïnvloeden. De overgrote meerderheid van de metadata is nu in het Dublin Core formaat. In sommige gevallen beschikken we ook over rijkere metadata als ESE of EAD.

De uitkomst van de data-analyse op een testset moeten de (kwaliteits)problemen in kaart brengen en wat dit betekent voor geocoderen en verrijken. Naast een analyse van de

kwaliteit van de metadata en de consequenties hiervan voor de ontsluiting. Daarnaast zal er advies worden gegeven voor praktische oplossingen.

Uitgangspunt van de pilot is dat geografische metadata in de volgende velden kan worden aangetroffen:

- titel (dc:title)
- beschrijving (dc:description)
- dekking/plaats/tijd (dc:coverage)
- onderwerp/trefwoorden (dc:subject)

Voor deze velden worden de opbrengsten en kwaliteit geanalyseerd. Ook wordt er per collectie bekeken wat de opbrengst is.

Voor het veld Titel kan eventueel Named Entity Recognition (NER) worden verkend. In coverage worden vaak vele geografische termen, locaties weergegeven. Deze zou mogelijk ook nog kunnen gebruikt voor geografische hierarische relaties (plaats-gemeente-provincie). In trefwoorden zijn soms preciezere locaties opgenomen dan in coverage. Mogelijk bevind zich in andere velden ook nog geografische metadata.

Werkwijze

Stap 1: extractie termen uit coverage

Eerst zijn geografische termen uit het veld:coverage opgenomen. Meest problematische daarbij was het reconstrueren van hiërarchie.

Die was op allerlei logische maar verschillende manieren opgenomen:

- kommagescheiden in één veld,
- in verschillende velden,
- in verschillende volgordes,
- met verschillende schrijfwijzes voor provincies, met opname gemeente,
- voorzien van aanduidingen als 'dorp:' of 'gemeente:',
- met opname streek of eiland of aangeduid met historische naam.

Regelmatig kwamen meerdere plaatsaanduidingen bij één record voor, zodat vaak onduidelijk was welke termen tot een hiërarchie behoorden en welke niet.

Vergelijk de volgende voorbeelden - wanneer is er sprake van hiërarchie en hoe vis je die eruit?

`haarlem`, `alkmaar`, `1930-1945`

`haarlem`, `nederland`, `noord-holland`

`haarlem`, `noord-holland`, `nederland`

`bloemendaal`, `haarlem`, `nederland`, `noord-holland`

`nederland`, `noord-holland`, `haarlem`, `haarlem`

`haarlem, westerbork, vught, auschwitz`

`indonesië`, `nederland`

`aek rioeng, indonesië, nederlands-indië`

`indonesië`, `nederlands-indië`, `pasoeroean`

`buitenzorg`, `java`, `kedoengbadak`, `nederlands-indië`

`bangkinan`, `nederlands-indië`, `sumatra`

`botosari, indonesië, nederlands-indië, poerwakarta`

`gemeente: Venray`,`dorp: Centrum`,`straat: Leeuwstraat`

Stap 2: extractie termen uit tekst

Van records zonder `coverage` zijn `title` en `description` tegen een postagger aangehouden om eigennamen te extraheren. Die eigennamen zijn vervolgens alleen als geografische term opgeslagen als ze door de reguliere expressie

```
(^| )(?i)(te|uit|in|op|bij|naar)(?-i) (de|het|den)?[ ]?Eigennaam
```

kwamen (waarbij `Eigennaam` vervangen werd door de door de postagger gevonden eigennaam).

Dit is gedaan met alle collecties die nu in Netwerk Oorlogsbronnen zijn opgenomen, behalve de collectie oorlogskranten. De aantallen van de collectie oorlogskranten (7.761.539 items) vertekenen het beeld. Er is een testset gekozen uit deze collectie van 190 in de oorlog

verschenen nummers van De Gelderlander. Deze selectie is ook gegeocodeerd, de werkwijze en bevindingen daarvan zijn in het document ‘gelderlander-documentatie.pdf’ beschreven.

Tabel 1: Resultaten zoeken in velden: coverage; title; description en subject

Totaal aantal records uit Oorlogsbronnen	800.911
Records met veld:coverage	176.912
Aantal termen gevonden in veld: coverage	214.667
Aantal termen gevonden in veld: Title of Description (tekstveld)	121.643
Aantal unieke termen*	26.260

* Een unieke term is bijv. Amsterdam, deze unieke term wordt gevonden in 11.184 records

Vooral de set OCLC (Bibliotheek NIOD) gaf problemen met deze aanpak, omdat veel bibliotheektitels in het Engels of Duits gesteld zijn.

Stap 3: geocoderen

De termen zijn eerst tegen de Historisch Geocoder gehouden, waarbij gebruik is gemaakt van de Erfgeoproxy (gebouwd boven op de Historische Geocoder, vereenvoudigt een aantal specifieke zoekacties).

Om de kans op een eenduidig resultaat te vergroten is eerst gezocht naar landen, vervolgens naar plaatsen, provincies, gemeenten, straten en tot slot adressen. Zodra er een resultaat was is steeds het zoeken gestopt, zodat bij Denemarken het land gevonden werd, en niet ook de gelijknamige plaats bij Slochteren.

Er is zoveel mogelijk gegeocodeert naar GeoNames URI’s, een keuze die is ingegeven door de wereldwijde dekking, de gebruiksvriendelijke API van GeoNames zelf en de brede toepassing van GeoNames URI’s in wetenschap en het erfgoedveld. Voor straten, adressen en gebouwen zijn BAG id's gebruikt. De BAG (Basisadministratie Adressen en Gebouwen) heeft nog geen URI’s (googlen op BAG URI brengt je bij plaszakjes voor dames en heren), maar de BAG id's zijn wel op te vragen bij de Historische Geocoder.

Bij één resultaat is het resultaat opgeslagen, bij meerdere resultaten is de term alleen gemarkeerd met 'multiple'. Dit betekent dat een term meerdere geocodes kan opleveren en het vergt (vaak handmatig) onderzoek om de juiste match te maken. Er ligt een Middelburg op Walcheren, maar ook een Middelburg in Zuid-Holland.

Termen die geen enkel resultaat gaven zijn vervolgens tegen de GeoNames API gehouden, waarbij een locatie als resultaat is bestempeld als de schrijfwijze exact overeen kwam. Daarbij moet wel een goede afweging gemaakt worden naar welke talen wel en niet gekeken wordt: ‘Brussel’ matcht niet met ‘Brussels’, maar wel met een Zuid-Afrikaanse boerderij genaamd ‘Brussel’.

Tabel 2: Unieke termen gegeocodeerd

	HGC	GeoNames	Totaal
één resultaat	9.710	858	10.568
meerdere resultaten	1151	957	2.108
geen resultaat (bevat veel *true negatives*)	189	13.370	13.584

De termen met één resultaat zijn onder te verdelen in de onderstaande typen. De lokatietypen met hoofdletter (behalve Point, waarbij geocoding niet nodig was) komen uit de Historische Geocoder, de overige uit GeoNames.

Tabel 3: Lokatietypen

count	type
3400	Place
3229	Point (coördinaten in `coverage`)
1938	Street
852	Address

454	Populated place
184	Country
78	Municipality
451	other (neighbourhood, hotel, museum, island, etc.)

Stap 4: extractie termen uit subject

Van records zonder `coverage` zijn, pas na het geocoderen, de trefwoorden uit het veld `subject` tegen de in de vorige stappen al geëxtraheerde termen gehouden. Trefwoorden die nog niet in die termen voorkwamen zijn dus niet opgenomen. Enerzijds omdat het tegen verschillende geocoders houden van alle nieuwe termen veel tijd zou kosten, anderzijds omdat de kans reëel was dat termen, die niet al eerder in coverage of tekst werden aangetroffen, vals positief zouden zijn. De uit tekst geëxtraheerde term ‘landbouw’ werd al als straat in Houten geïnterpreteerd - in hetzelfde wijkje liggen ook de straten ‘Tuinbouw’, ‘Stedebouw’, ‘Wegenbouw’ en ‘Scheepsbouw’.

Tabel 3: aantallen records met term uit ‘subject’

Aantal records met term uit subject	411.565
waarbij term 1 resultaat heeft	248.169
waarbij term geen resultaat heeft	208.598
waarbij term meerdere resultaten heeft	49.791

Resultaten en analyse

De resultaten verschillen aanzienlijk per dataset, zoals onderstaande tabel laat zien. De 100 procent score van de collectie OORLOGSMONUMENTEN torent hoog uit boven de 8 procent van de collectie OCLC (NIOD bibliotheek).

Records met een coverage veld scoren, zoals verwacht, aanzienlijk beter dan wanneer we geografische termen in tekst moeten zien te vinden . Opvallend is wel dat termen uit coverage lang niet altijd eenduidig gegeocodeerd kunnen worden - vergelijk de 120 duizend records met coverage veld van BBWO2, waarvan er maar 80 duizend gegeocodeerd zijn.

term uit coverage	**resultaat uit coverage**	**idem, % van totaal**	**term uit text**	**resultaat uit text**	**idem, % van totaal**	**term uit subject**	**resultaat uit subject**	**idem, % van totaal**
ARCHIEVEN NIOD archieven en collecties							151988 records	
0	0	0 %	32653	21735	14 %	0	0	0 %
ARCHIEVENWO2 Archieven WO2							48362 records	
2714	2617	5 %	7587	5141	11 %	19072	0	0 %
BBNA Beeldbank Nationaal Archief							24920 records	
19538	15101	61 %	779	361	1 %	3991	212	1 %

BBWO2 Beeldbank WO2							132976 records	
119650	99492	75 %	2236	1303	1 %	3601	113	0 %
DANS DANS-KNAW - diverse collecties							240 records	
237	209	87 %	0	0	0 %	1	0	0 %
DIGCOL Digitale collecties							33418 records	
1909	1337	4 %	11352	5207	16 %	15045	515	2 %
DIMCON Digitale Museale Collectie Nederland							23476 records	
20907	17509	75 %	851	527	2 %	1346	778	3 %
GETUIGENVERHALEN Getuigenverhalen							655 records	
587	470	72 %	14	14	2 %	19	0	0 %
GVNEVDO01 Geheugen van Nederland - Oorlogsdagboeken							802 records	
0	0	0 %	416	397	50 %	782	299	37 %

GVNEVDO02 Geheugen van Nederland - Propagandadrukwerk WO II							2935 records	
0	0	0 %	290	155	5 %	2830	1211	41 %
GVNEVDO03 Geheugen van Nederland - Verzetsliteratuur							6151 records	
0	0	0 %	234	109	2 %	6109	355	6 %
GVNMUSE01 Geheugen van Nederland - Kamptekeningen uit bezet Nederlands-Indië (1942-1945)							4938 records	
0	0	0 %	833	218	4 %	4878	3228	65 %
GVNNIOD02 Geheugen van Nederland - Illegale pamfletten en brochures							1234 records	
0	0	0 %	164	92	7 %	1234	1197	97 %
IIO_LEGER Fotocollectie Dienst voor legercontacten Indonesië							7048 records	
7014	3252	46 %	0	0	0 %	0	0	0 %
IIO_ONAFHANKELIJK Indonesië onafhankelijk 1947-1951							4580 records	
0	0	0 %	239	161	4 %	0	0	0 %

IPNV Interviewproject Nederlandse Veteranen							101 records	
79	78	77 %	22	18	18 %	22	16	16 %
KITLV Beeldbank KITLV							15189 records	
0	0	0 %	9270	3576	24 %	15145	15110	99 %
MFORCE-MEDIA WO2 in Muziek							72 records	
0	0	0 %	39	25	35 %	0	0	0 %
NATIONAAL-ARCHIEF Archieven Nationaal Archief							76261 records	
0	0	0 %	16576	10555	14 %	0	0	0 %
OCLC NIOD bibliotheek							62157 records	
0	0	0 %	9809	4836	8 %	57157	48303	78 %
OIB Oorlog in Blik							742 records	
455	404	54 %	172	156	21 %	0	0	0 %

OORLOGSGRAVEN Oorlogsgraven							169353 records	
0	0	0 %	0	0	0 %	38158	56	0 %
OORLOGSMONUMENTEN Oorlogsmonumenten							3725 records	
3725	3724	100 %	0	0	0 %	0	0	0 %
RAL Regioneel Archief Leiden							29588 records	
0	0	0 %	1751	492	2 %	0	0	0 %

Vals positieven

Maar hoe goed zijn de resultaten eigenlijk? We hebben een steekproef gedaan van tweehonderd random gekozen uit `coverage` afkomstige termen en een zelfde steekproef voor termen afkomstig uit tekstvelden.

Van die tweehonderd uit `coverage` afkomstige termen was 0.5% verkeerd gegeocodeerd - `Atlantische Oceaan` werd beschouwd als straat in Naaldwijk. Creatieve straatnaamcommissies zorgen overigens voor meer problemen - buiten de steekproef kwamen we in Purmerend een wijkje tegen met straten als `Bali`, `Kalimantan`, `Sulawesi` en `Borneo`.

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 (25 van de 100 vals positief) of termen die weliswaar in tekst zijn aangetroffen, maar die soms ook in coverage voorkomen (12 van de 100 vals positief).

Ter illustratie: 'Groothandelaren in Glas' heeft niets te maken met de Oostenrijkse plaats `Glas`, met de 'Lofzang tot God in Sion' had men vast niet het Rijswijkse Sionbuurtje op 't oog en het 'Kriegsschiff in Not' zal niet in het Oostenrijkse `Not` voor anker hebben gelegen.

Ook hier nemen straatnaamcommissies je soms flink grazen: in Aalten ligt `Het Verzet` iets ten oosten van `Bevrijding`, `Spitfire` is een straat in Nootdorp en `Anne Frank` een straat in Bunschoten-Spakenburg (om de hoek bij `Grebbelinie`).

Op basis van deze resultaten kan je stellen dat je van gegeocodeerde termen uit 'coverage' vrij zeker kunt zijn, maar dat je met uit tekst geëxtraheerde termen op moet passen - bijna een vijfde van het totaal is fout gegeocodeerd. Dit wil overigens niet zeggen dat het met een vijfde van de *records* mis gaat - 'Glas' kwam maar in één record voor, terwijl correct gegeocodeerde termen als 'Japan', 'Berlijn' en 'Arnhem' elk in honderden records zijn gevonden.

Vals positieven elimineren

Er is wel het één en ander te bedenken waar de kwaliteit van de data mee verbeterd kan worden. Zo zal alleen al het kritisch kijken naar gevonden straatnamen waar geen 'straat', 'weg', 'plein', etc. in voorkomt al een hoop problemen oplossen. Ook gebieden of landen die je niet zo snel binnen je dataset verwacht zou je aan een nader onderzoek kunnen onderwerpen. Zo geocodeerden we pittoreske plaatsjes in de Verenigde Staten als 'Uniform', 'February', 'August', 'Volt', 'Exile', 'Library', 'Channel', 'Lord', 'Brief', 'Social', 'Sector' en 'Axis'.

In het geval van oorlogsbronnen konden we ook kijken welke termen niet in 'coverage' gevonden waren, en wel - veel - in tekst. In de toptien waren alleen de - Engelstalige - termen 'Berlin' en 'Germany' goed gegeocodeerd. De overige termen waren, van veel naar meest voorkomend: 'Rijk' (plaats in België), 'Ridderzaal' (straat in Eindhoven), 'augustus' (straat in Wijk bij Duurstede), 'Straat' (plaats bij Roermond), 'november' (straat in Heerhugowaard), 'Markt' (plaats in Gelderland), 'Zoom' (plaats bij Nunspeet) en met 515 records afgetekend op de eerste plaats: 'Engels' (plaats in Noorwegen).

Voorbeelden als hierboven gegeven hebben we ter illustratie op een Vals Positieven Kaart gezet.

Vals negatieven

Het zal opgevallen zijn dat een zeer groot aantal termen (13.584) geen resultaat heeft opgeleverd. Dit hoeft geen tekortkoming te zijn, want er zijn met Named Entity Recognition veel termen uit tekstvelden geplukt die niets met geografie van doen hebben. Dat `Roode Leger`, `Houthakkers`, `Afdeling Algemeen Secretariaat`, `Reconstructie Bijeenkomst`, `Werkgeversverklaring|Voorbeeld`, `Feldurteile`, `Alarmbereitschaft` en `Ausland` niets opgeleverd hebben is terecht, al zou je van de laatste term kunnen zeggen dat die wel een (niet heel specifieke) geografie aanduidt.

Een steekproef van honderd random gekozen uit tekst afkomstige resultaatloze termen leverde het volgende op:

51	geen geografische term	Zie de voorbeelden 'Roode Leger' en verder hierboven

13	gebouw, kamp, etc.	Wel lokatie, niet in geocoders te vinden, vaak ook historische naam: 'Haagse Koninklijke Schouwburg', 'Benthienkazerne', 'Wang Po', 'R.K. U.L.O.', 'Philipsfabrieken'.
12	wel geografische term, spelling afwijkend	'Mildwolda' i.p.v. Midwolda, 'Stassfurt' ipv Staßfurt, 'Oesterreich' i.p.v. Österreich, samentrekkingen als 'Rotterdam-Hillegersberg' en 'Tel Aviv-Jaffa', regio-aanduidingen als 'Eastern-Indonesia'
10	historische spelling of naam	'Denenmarken', 'Rijksweg Amsterdam-Velsen', veel Indonesische namen als 'Tjipalat' (Cipalat), 'Poerbolinggo' (Purbalingga), 'Djokja' (Jogjakarta)
9	organisatie, soms geografisch te duiden	'Mobilisatiebureau', 'H.Landstichting', 'Djokjasche vliegclub', ook kranten als 'N.R.C. Handelsblad' en 'Leeuwarder Koerier'
6	bijvoegelijk gebruikte geografische term	'Loosdrechtse', 'Mexicaanse', 'Zaanse', 'Haagse'

Op basis van deze bescheiden steekproef zou je kunnen stellen dat 35 tot 45 procent van deze resultaatloze termen vals negatief is.

Kijk je naar records in plaats van naar termen, dan is dat percentage lager - gesorteerd op aantal records zitten er tussen de 30 meestvoorkomende uit tekst geëxtraheerde resultaatloze termen maar zeven die met enige wil geografisch genoemd zouden kunnen worden (waarbij ‘Zondagmiddagcabaret’ in twee keer zoveel records blijkt voor te komen als ‘Europe’).

Zeker is dat er bij de extractie van termen uit tekst ook veel geografische aanduidingen over het hoofd zijn gezien. Dat zou het percentage weer hoger maken. Het blijft gissen, kortom, maar dat pakweg de helft niet wordt opgepikt of niet eenduidig wordt gegeocodeerd lijkt aannemelijk.

Vals negatieven alsnog oplossen

Het zou goed zijn te kijken naar een lijst historische Indonesische namen. Voorts zouden een lijst kampen ('Wang Po'), andere oorlogsgerelateerde geografische namen (‘Hitler-Deutschland’) of een referentielijst verdwenen bouwwerken ('Bentheinkazerne') ook helpen.

Geautomatiseerd zou een enkele term nog wel op te lossen zijn ('ss/ß' en 'oe/ö', samentrekkingen uit elkaar halen), maar veel termen ontberen context ('Zaanse wethouder' of 'Zaanse mosterd', welk 'mobilisatiebureau'?).

Handmatig sla je snel een flinke slag door de meestvoorkomende louter uit tekst geëxtraheerde resultaatloze termen even langs te lopen. Door alleen ‘den Niederlanden’ en ‘Deutschland’ op te lossen geocodeer je in één klap 1994 records. Uiteindelijk hebben we iets van zeventig termen handmatig van GeoNames URI’s voorzien, en daarmee ongeveer twintigduizend records alsnog gegeocodeerd.

De enige manier om een honderd procent score te naderen lijkt vooralsnog alleen haalbaar door een mens elk record afzonderlijk te laten beschrijven.

Gegeocodeerd, en dan?

Geografisch zoeken - mogelijkheden

Dat een kaart je zoekresultaten inzichtelijker maakt zal niemand ontkennen - je krijgt meteen een goed beeld van de geografische spreiding van hetgeen je naar zocht.
Geometrieën (puntjes, lijnen of polygonen) geven je de mogelijkheid te zoeken op nabijheid, binnen een bounding box (een rechthoek tussen twee schuin tegenover elkaar liggende coördinaten) of een polygoon (een veelhoek die je bijvoorbeeld de mogelijkheid geeft te zoeken binnen het gebied ‘De Veluwe’).

figuur 1: inzicht in aantallen records per woonplaats

Een goed gestandaardiseerde geografische aanduiding brengt behalve geometrie ook hiërarchie binnen handbereik. Van een GeoName of BAG id kan je eenvoudig achterhalen in

welke plaats, gemeente, provincie of welk land het ligt. Sla je deze hiërarchische gegevens op, dan kan je dus goed zoeken op ‘verzetskranten gelderland’ - ook als die verzetskranten oorspronkelijk alleen met plaatsnamen getagd waren. Dit is ook te gebruiken voor exports.

Om dit alles mogelijk te maken moeten de eenduidige resultaten die het geocoderen heeft opgeleverd opgenomen worden in de Elastic Search index van Netwerk Oorlogsbronnen. Binnen deze pilot zullen NDJSON bestanden gemaakt worden waarmee Trifork dit eenvoudig kan doen. NDJSON staat voor ‘newline delimited json’, een formaat waarin elke regel een JSON array bevat, met in dit geval de NIOD identifier, geometrie en hiërarchie: plaats, gemeente, provincie, land.

Zijn de resultaten in de Elastic Search index opgenomen, dan is het aanpassen van de tekstuele zoekinterface relatief weinig werk. De nieuw geïndexeerde velden zullen binnen de bestaande API op Elastic Search toegankelijk komen. De opgeslagen hiërarchie komt vooral hierbij van pas. Een kaartinterface zal wat meer inspanning kosten, maar ook die is te overzien. Sowieso is het aan te bevelen een prototype van zo’n kaartinterface te maken, al was het maar om intern de waarde van het geocoderingsproces verder te kunnen evalueren.

Verrijkte data thuisbrengen

Het is een goed idee enige moeite te doen verrijkingen in het collectie beheer systeem van de data-eigenaar zelf te krijgen. Daar lopen de verrijkingen de minste kans verloren te gaan.

Staan de verrijkingen ergens bij een aggregator op een server, dan is het met de gegevens gedaan zodra de aggregator ermee ophoudt. Of de gegevens worden na een aantal jaren door de aggregator zelf terzijde geschoven omdat de kwaliteit niet optimaal is. De kwaliteit zal in ieder geval niet verbeteren, want het is niet mogelijk de gegevens te editen.

Ook niet denkbeeldig is dat de data-eigenaar zijn 'permalinks' of 'persistent' identifiers (als die er al waren) wijzigt en de koppelingen in de verrijkingen verwijzen naar een niet meer terug te vinden object. In de praktijk gaan persistent identifiers vaak niet langer mee dan de periode waarin een data-eigenaar een bepaald collectiebeheersysteem gebruikt.

Voor erfgoedinstellingen is het van groot belang is de verrijkte data in de eigen collectiebeheersystemen wordt opgenomen. Met deze verrijkingen kan ook binnen eigen systemen de “waar-”vraag beter beantwoord worden, hetgeen zowel geografisch zoeken als een geografische presentatie op de kaart dichterbij brengt. Het opnemen in de eigen systemen vereist alleen een juiste koppeling op uniek identificatienummer en mogelijk (extra) velden waarin de geografische verrijkingen terecht kunnen komen. Als de erfgoedinstelling redactie voert zullen de verrijkingen nog beter worden en zal de “waar”-vraag in de toekomst, zowel binnen de eigen systemen als op andere aggregerende platformen (Europeana, DimCon), nog beter beantwoord kunnen worden. In het eigen collectiebeheersysteem schrijft en schaaft de data-eigenaar immers al regelmatig aan zijn collectiedata.

Een argument dat soms tegen teruglevering gebruikt wordt is dat het gebruikte collectiebeheersysteem er 'niet klaar voor is'. Inderdaad zou het fijn zijn als die softwarepakketten online koppelingen met veelgebruikte thesauri zouden faciliteren, maar

welbeschouwd is er meestal wel een weg als er een wil is. Voor het opslaan van een URI is niet meer dan een tekstveld nodig.

Voor het NIOD betekent dit dat de geocoderingen op haar collecties (NIOD archieven en collecties, Beeldbank WO2 en NIOD bibliotheek) in haar bestaande collectiebeheersystemen zullen moeten worden opgenomen. Bekenen moet worden of hier aanpassingen aan de systemen door de verschillende leveranciers voor nodig zijn.

De verrijkingen zouden voor de leverende instellingen klaargezet moeten worden, zodat die de verrijkingen in kunnen lezen in het eigen systeem. Een korte rondvraag langs enkele leverende instellingen zou de vraag moeten beantwoorden hoe dat het handigst te doen is. Via oai-pmh is een optie, maar misschien hebben instellingen liever een csv-bestand. Onderzoek moet uitwijzen of de verrijkingen in alle gevallen überhaupt nog te koppelen zijn (komen de identifiers in Oorlogsbronnen nog overeen met die in de collecties zelf?).

Geocoderen tijdens het aggregeren

Bekijken hoe records tijdens het aggregeren gegeocodeerd kunnen worden was deel van de opdracht van deze pilot. Het moge duidelijk zijn dat dit een lastig proces is met veel kans op zowel vals positieven als vals negatieven. In de pilot zijn in de scripts, bijvoorbeeld om hiërarchie af te leiden uit het ‘coverage’ veld, regelmatig op specifieke datasets toegespitste stukjes code geschreven. Tijdens het proces kan het nodig zijn om op basis van de resultaten de code enigszins aan te passen. Geocoderen geeft, kortom, betere resultaten als het geen louter automatisch proces is.
Praktisch is het aanroepen van API’s een duur proces - het aggregeren duurt zo een factor tien langer dan zonder. En de GeoNames limiet van tweeduizend aanroepen per uur voor gratis gebruik speelt ook mee.

Wil je ondanks dat het geocoderen toch automatisch tijdens het aggregeren laten geschieden, dan is het logisch je te beperken tot termen uit ‘coverage’ (termen uit tekst geven, zonder enige controle, teveel vals positieven). Je zou je verder kunnen beperken (en de kans op vals positieven kunnen verkleinen) door termen niet tegen een api aan te houden, maar tegen een lijst van eerder gegeocodeerde (en enigszins gecontroleerde) termen.

Als deze pilot één ding heeft duidelijk gemaakt, dan is het dat gebruik van URI’s of id’s van geografische thesauri als GeoNames, TGN of BAG veel problemen oplossen - de eenduidigheid die ze met zich meebrengen reduceert de kans op vals positieven in ieder geval tot vrijwel nul. Een script dat bij het aantreffen van zo’n URI gegevens als geometrie en hiërarchie ophaalt is dan ook redelijk eenvoudig te schrijven.

De aanbeveling is dan ook om zo’n script in het aggregatieproces op te nemen. Op tijd die nodig is om API’s aan te roepen kan bespaard worden door intern een database aan te leggen van URI’s met benodigde informatie - dan hoef je niet steeds opnieuw geometrie en hiërarchie van ‘Amsterdam’ op te halen. Het faciliteren en ‘belonen’ van instellingen die

URI’s gebruiken zal uiteindelijk juist de aggregator ten goede komen, omdat aangeleverde data uiteindelijk eenduidiger zal zijn.

De koninklijke weg om data van geografische verrijkingen te voorzien is: gecombineerd scriptmatig / handmatig geocoderen als in deze pilot, de resultaten opnemen in het collectiebeheersysteem van leverende instelling, bij export van zo’n collectie URI’s opnemen in ‘coverage’, bij inlezen van collectie door Oorlogsbronnen op basis van URI naam, hiërarchie en geometrie ophalen.

Conclusies en aanbevelingen

GeoNames is de handigste thesaurus gebleken om plaatsen, provincies, landen (en typen als water, eiland, museum, etc) mee te benoemen.

Geografische thesauri verbeteren helpt jezelf en anderen. We hebben een aantal historische namen (‘Nederlands-Indië’, ‘Sovjet-Unie’, ‘Joegoslavië’, ‘Oranjehotel’) en een aantal kampen (‘Kampong Makassar’, ‘Lampersari’, ‘Kamp Westerbork’) aan GeoNames toegevoegd.

Het NIOD zou kunnen overwegen de intern gebruikte lijst met kampen, etc. te publiceren, liefst als linked data. Daarbij kunnen o.a. links naar bestaande of aan te maken GeoNames items opgenomen worden. Dit vanuit het idee dat het NIOD niet alleen de aangewezen partij is om oorlogsgerelateerde collectiemetadata centraal te ontsluiten, maar dat datzelfde geldt voor oorlogsgerelateerde terminologie.

De BAG is de beste (en eigenlijk ook de enige) thesaurus gebleken om (huidige Nederlandse) adressen en gebouwen te benoemen.

Termen uit Coverage leveren vrijwel geen false positives op, maar een kwart tot een derde van de termen is niet in één keer eenduidig te geocoderen.

Met NER verkregen termen uit tekstvelden komen we op 10-20% false positives. Met semi-automatische processen is dat percentage tot onder de 10% te brengen.

In de hele keten (aggregatie, collectiebeheersysteem, data-ontsluiting) zou gebruik van URI’s mogelijk gemaakt moeten worden.

Verrijkingen die niet in het collectiebeheersysteem, maar alleen bij een aggregator leven zijn beperkt houdbaar.

De verrijkingen moeten aan de leverende instellingen worden aangeboden.

Het NIOD zou de verrijkingen in ieder geval in haar eigen collectiebeheersystemen (Bibliotheek, Archief en Beeldbank) op moeten nemen.