Van documenten naar data

Archieven en het semantisch web

Edwin Klijn ■

Achttien jaar terug in de tijd: in mei 2001 publiceerde Tim Berners Lee – de man die eerder aan de wieg stond van het World Wide Web – samen met James Hendler en Ora Lassila een opzienbarend artikel in het tijdschrift Scientific American over wat zij het 'Semantic Web' noemden.[1] Daarmee beschreven zij de transformatie van een 'web of documents' naar een 'web of data'.

Linked Open Data is aan een opmars bezig. Er worden steeds meer inhoudelijke verbindingen gemaakt op basis van de inhoud van informatiebronnen. Voor archiefinstellingen biedt het Semantisch Web dan ook nieuwe mogelijkheden om collecties voor een wereldwijd publiek toegankelijk te maken en bronnen te contextualiseren.

Hoe word je onderdeel van het Semantisch Web? Het begin is een gedigitaliseerd archief, maar dan ben je er nog lang niet. Een scan van een bladzijde met tekst is voor een computer niet veel anders dan een vakantiefoto. Om de bladzijde digitaal doorzoekbaar te maken, moeten de letters en woorden door software worden omgezet naar machine-leesbare tekst. Om zoektoegangen te maken is 'verrijking' van deze tekst nodig. Ook voor deze stap is software voorhanden.
Hoe je zo efficiënt mogelijk van papieren archief naar een semantisch verrijkt digitaal corpus kunt komen, wordt onderzocht in het TRIADO-project (Tribunaalarchieven als Digitale Onderzoeksfaciliteit). In TRIADO doen het Nationaal Archief, Huygens ING, NIOD en Netwerk Oorlogsbronnen (coördinator) onderzoek naar de mogelijkheden om met nieuwe digitale technologieën archiefdocumenten machine-leesbaar te maken en vervolgens te verrijken, zodat er toegangen op persoonsnamen, locaties, organisaties, datums en typen document kunnen worden gebouwd. Als testcase is een steekproef van 13,8 strekkende meter (167.197 bladzijden) geselecteerd uit het meest geraadpleegde Tweede Wereldoorlog-archief in Nederland: het Centraal Archief voor Bijzondere Rechtspleging (CABR). Dit archief van vier strekkende kilometer omvat de justitiële stukken van circa 300.000 personen die na de oorlog verdacht werden van collaboratie met de Duitse bezetter. Het bestaat uit processen-verbaal, besluiten, inlichtingenformulieren, lidmaatschapskaarten, foto's en allerlei archiefmateriaal dat als bewijsstukken is toegevoegd aan de dossiers. Het gaat overwegend om typoscripten.

'Well-defined meaning'

'The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The first steps in weaving the Semantic Web into the structure of the existing Web are already under way. In the near future, these developments will usher in significant new functionality as machines become much better able to process and 'understand' the data that they merely display at present.'
Bron: Scientific American (May 2001).[1]

Steekproef

In TRIADO is de steekproef met standaard OCR-software (Optical Character Recognition) machine-leesbaar gemaakt. Deze ruwe data vormen de basis voor full-tekst zoekfuncties en de ontwikkeling van zoektoegangen. Vervolgens is de kwaliteit van de machine-leesbare tekst gemeten. Met ABBYY FineReader werd zonder enige training of voorbewerking een 'word error rate' (WER = percentage woorden dat incorrect door de software is

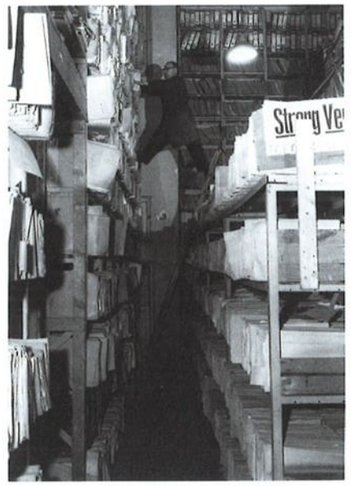
NSB-lidmaatschapskaart voor 1934 van stamboeknummer 1 Anton Mussert (WikiCommons, publiek domein).

omgezet) gemeten van 15. Een verre van perfecte transcriptie, maar nog altijd voldoende om te doorzoeken en nadere toegangen te ontwikkelen. In de test-website die voor dit project is ontwikkeld, leverde een eenvoudige zoekactie op de term 'Oostfront' al 619 resultaten op.
Vervolgens is er onderzocht in hoeverre er nadere toegangen door de computer kunnen worden gegenereerd. 'Named entity recognition'-software is gebruikt om namen van personen, organisaties, producten en locaties automatisch te herkennen. Dit bleek heel moeilijk te zijn, onder andere vanwege de Duitse woorden in de tekst en de gewoonte in de jaren veertig om functies met hoofdletters te schrijven. Het matchen aan bestaande databestanden bleek een valide alternatieve strategie. Met behulp van het bestand van de Nationale Database Vervolgings Slachtoffers (NDVS) konden in de 13,8 meter testmateriaal al enkele slachtoffers geïdentificeerd worden.
In TRIADO is ook geëxperimenteerd met het automatisch herkennen van documenttypen. Voor 28 vooraf bepaalde categorieën (processen-verbaal, besluiten et cetera) zijn voorbeelden verzameld. De computer is op basis van deze data zichzelf gaan leren documenten te classificeren. Aan het einde van de test was de foutmarge 20%. Dit lijkt hoog, maar getypte besluiten en getypte processen-verbaal zijn ook voor het menselijk oog moeilijk te onderscheiden. De beste scores werden gehaald met documenten met een vaste opmaak, zoals bijvoorbeeld NSB-lidmaatschapskaarten.

Andere nieuwe vormen van archiefontsluiting deden zich voor bij de experimenten met zogenoemde 'topic modelling'. Deze technologie analyseert de machine-leesbare tekst en geeft op basis daarvan een korte karakterisering van een of meer documenten in steekwoorden. Neem bijvoorbeeld:

> *Topic #17: groningen landwachters <naam> <naam> gearresteerd slochteren <naam> siddeburen ondergedoken woning landwacht <naam> arrestatie <naam> overgebracht onderduikers west gemeente <naam> schildwolde duitsland onderduiker huiszoeking getuige landbouwer personen boerderij wonende <naam> <naam>*

Dit betreft een proces-verbaal dat handelt over een zaak in Groningen waarbij leden van de Landwacht – een hulppolitie bestaande uit NSB'ers – ervan verdacht werden betrokken te zijn geweest bij het oppakken van enkele onderduikers. Topic modelling bleek in de tests in TRIADO een handige manier om inhoudelijke informatie te geven over een of meerdere documenten.



Zoeken door CABR-dossiers. Bron: fotoalbum 'Centraal Archievendepot Justitie', Nationaal Archief.

Uitermate geschikt

Het TRIADO-project is nog niet afgerond. In de laatste fase tot juli van dit jaar verkent NIOD-onderzoeker Ismee Tames aan de hand van enkele use-cases de potentie van het digitale corpus voor nieuw wetenschappelijk onderzoek. Wat betreft het ontsluiten van archiefcollecties luidt de conclusie dat automatische tekstherkenning, autoclassificatie, dataverrijking door matching aan bestaande databestanden en topic modelling ondanks alle onvolkomenheden nu al uitermate geschikt zijn om grote hoeveelheden gedigitaliseerd archiefmateriaal (semi-)automatisch toegankelijk te maken.
Niet alleen TRIADO geeft aanleiding tot optimisme: met artificiële intelligentie en machine learning worden er in het project Transkribus inmiddels indrukwekkende scores gehaald met handgeschreven bronnen. Voor zeventiende-eeuwse Amsterdamse notariële akten worden met een bescheiden hoeveelheid handmatig ingevoerde trainingsdata al 'character error rates' bereikt van rond de 5.[2] Machine learning wordt inmiddels ook ingezet om softwarematig structuren in een tekst te herkennen (bijvoorbeeld in het geval van een kasboek) en deze om te zetten naar spreadsheets. En wat te denken van het automatisch herkennen van gezichten op foto's of handtekeningen in geschreven correspondentie.
Voor Netwerk Oorlogsbronnen – als samenwerkingsverband dat tot doel heeft 'de collectie WO2 Nederland' digitaal beter zichtbaar te maken voor het publiek – is het vooruitzicht dat archiefcollecties tot op documentniveau machine-leesbaar worden gemaakt, verheugend. Waar de volgende uitdaging ligt, is dat deze data gekoppeld worden aan standaard referentiedata voor personen, plaatsen, datums, organisaties, gebeurtenissen et cetera. Neem bijvoorbeeld het volgende fragment:

> *Naam en voornamen: Jansen, Paul Geboortep!aats en-datum: Zaandam, 29 October 1897 Echtgenoote van / Beroep: voorheen agent van Politie, t Laatste woonptaats en adres: Kanaalstraat 25 II Amsterdam Persoonsbewijs-no.:z 2 01239 afgegeven te Leeuwarden Nationaiiteit (evt. vroeger) Nederlander die ervan verdacht wordt: joodsche personen in macht van den vijand te hebben gebracht, terwijl hij in dienst was van de S.D.*

Verrijkt met referenties naar Wikidata, Basisregistratie Adressen en Gebouwen en de WO2-thesaurus zou dit er idealiter als volgt uitzien: >>

Bronnen over de Sicherheitsdienst en contextinformatie op oorlogsbronnen.nl.

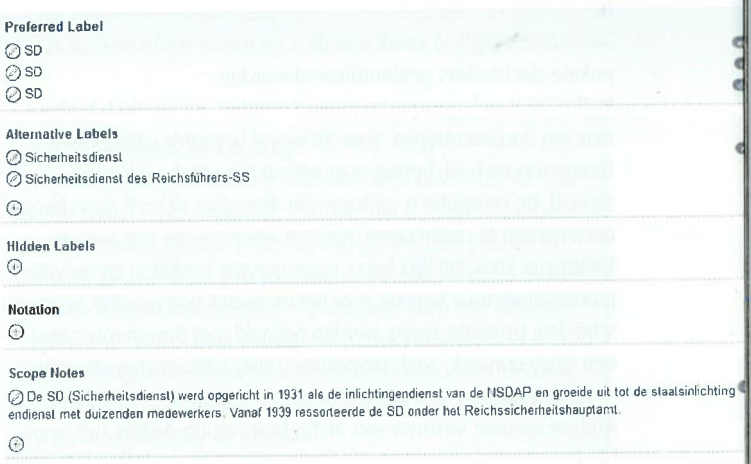
>> *Naam en voornamen: Jansen, Paul [https://data.niod.nl/WO2_biografieen/Paul-Jansen] Geboortep!aats en-datum: Zaandam [https://www.wikidata.org/wiki/Q211260], 29 October 1897 [dateformat:dd month yyyy] Echtgenoote van / Beroep: voorheen agent van Politie [https://data.niod.nl/WO2_Thesaurus/2715], t Laatste woonptaats en adres: Kanaalstraat 25 II Amsterdam[BA GID:0363200000159058] Persoonsbewijs [https://data.niod.nl/WO2_Thesaurus/1731]-no.:z 2 01239 afgegeven te Leeuwarden [https://www.wikidata.org/wiki/Q2311189] Nationaiiteit (evt. vroeger) Nederlander [https://www.wikidata.org/wiki/Q55] die ervan verdacht wordt: joodsche personen in macht van den vijand te hebben gebracht, terwijl hij in dienst was van de S.D.[https://data.niod.nl/WO2_Thesaurus/corporaties/4673]*

Verrijkingen maken het mogelijk meer te weten te komen over het leven van Paul Jansen, de Sicherheitsdienst en het pand aan de Kanaalstraat 25 te Amsterdam. Wat lastig is voor software (en vaak ook voor mensen!) is hoe om te gaan met ambivalente termen. Hoe kan de computer weten dat S.D. refereert aan Sicherheitsdienst en niet aan het vrachtwagenmerk SD? Welke Paul Jansen wordt hier bedoeld? In Netwerk Oorlogsbronnen wordt zo veel mogelijk informatie verzameld en aan de computer gevoerd, zodat deze slimmere matches kan maken.

Thesaurus

Op dit moment bouwt Netwerk Oorlogsbronnen aan een WO2-thesaurus met 'encyclopedische' data over concepten (verzet, collaboratie et cetera), personen, organisaties en locaties.[3] Deze terminologiebron, publiekelijk beschikbaar als Linked Open Data, is een uitstekende hulp om de logische samenhang tussen dit soort data vast te leggen op een manier dat ook de computer ermee uit de voeten kan. Een thesaurus kan achtergrondinformatie geven, relaties leggen en is toegerust op meertaligheid. In de WO2-thesaurus zijn ook koppelingen naar Wikidata opgenomen. De informatie uit de thesaurus kan worden gebruikt bij de presentatie, maar ook om de zoekfunctie te verbeteren, zoals nu op de site van Netwerk Oorlogsbronnen het geval is.

Conclusie van TRIADO en meer in het algemeen is dat computertechnologie grote potentie heeft om van documenten naar data te komen. Door data vervolgens te koppelen aan terminologiebronnen kan de brug worden geslagen naar het Semantisch Web. Voor nu is het de kunst om manieren te vinden hoe om te gaan met imperfectie. En collecties te ontsluiten als middel om andere collecties te ontsluiten. Het internet is een studiezaal die altijd open is. Wanneer kennisinstellingen hun expertise 'dataficeren' in terminologiebronnen en archiefinstellingen hun collecties machine-leesbaar, duurzaam, gestandaardiseerd en als open data aanbieden, kan gezamenlijk de eigen 'niche' op het web worden geclaimd. ■



De term 'SD' in de WO2-thesaurus.

Noten

1 T. Berners-Lee, James Hendler and Ora Lassila, 'The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities', *Scientific American* (May 2001). Zie: https://bit.ly/1ES18qe

2 Zie: https://bit.ly/2vTqZDD

3 Zie: https://data.niod.nl/WO2_Thesaurus.html

Alle voorbeelden zijn wegens privacyoverwegingen gesimuleerd. Lees het TRIADO-rapport op: www.oorlogsbronnen.nl/rapport-triado-verrijkingsfase. In september zal er een eindconferentie rondom TRIADO worden georganiseerd.

Edwin Klijn ■ programmamanager Netwerk Oorlogsbronnen en projectleider TRIADO.