Enorme stap om archieven voorwaarts toegankelijk te maken

Het Centraal Archief Bijzondere Rechtspleging telt ruim 300.000 dossiers van personen die na de Tweede Wereldoorlog zijn onderzocht omdat ze werden verdacht van 'collaboratie'. In het project Tribunaalarchieven als Digitale Onderzoeksfaciliteit (TRIADO) worden de mogelijkheden om dit archief digitaal te ontsluiten onderzocht.

Het centraal kaartsysteem van het Centraal Archief Bijzondere Rechtspleging

Bron: fotoalbum 'Centraal Archiefdepot Justitie'; Nationaal Archief

Edwin Klijn
Projectleider van TRIADO en programmamanager van Netwerk Oorlogsbronnen

Historische archieven bevatten een rijkdom aan informatie over personen, organisaties, gebeurtenissen, locaties en heel veel meer. De toegang tot het materiaal beperkt zich doorgaans tot enkele algemene aanduidingen: 'Correspondentie van de NSB in de periode van januari tot december 1934' of 'Ingekomen en uitgegane brieven van de Binnenlandse Strijdkrachten, 1945'. Volgens onderzoek van het ENUMERATE-project uit 2017 – uitgevoerd door Digitaal Erfgoed Nederland – is circa 11 procent van alle Nederlandse archieven gedigitaliseerd. Een nog kleiner percentage is daadwerkelijk digitaal te doorzoeken op elk woord in de tekst. Eenieder die wil weten wat er daadwerkelijk aan informatie in een archief is te vinden, moet de stukken opvragen en er handmatig doorheen bladeren.

Er is allerlei computertechnologie voorhanden om grote hoeveelheden documenten – en vooral de informatie in deze stukken – digitaal doorzoekbaar te maken tot op bladzijdeniveau. Het gaat hierbij met name om software die de inhoud van tekstuele bronnen – getypte of gedrukte documenten, maar ook handgeschreven stukken – converteert naar machineleesbare tekst. Daarnaast kan software getraind worden om allerlei 'verrijkingen' toe te passen. Hiermee kunnen in een webomgeving 'filters' worden gemaakt waarmee gebruikers hun zoekresultaten kunnen verfijnen.

In 2016 startten het Nationaal Archief, Huygens ING, NIOD en Netwerk Oorlogsbronnen (NOB) het project Tribunaalarchieven als Digitale Onderzoeksfaciliteit (TRIADO). De bedoeling was om op basis van een steekproef uit het Centraal Archief Bijzondere Rechtspleging (CABR) de mogelijkheden van digitale ontsluiting te verkennen door te experimenteren met 'proven technology', oftewel bestaande software, zowel open source als commercieel.

Twee pakketten getest

Een scan van een tekstdocument is voor een computer niet veel anders dan een vakantiefoto. Pas wanneer tekstherkenningssoftware de pixels transformeert naar letters en woorden, kan de tekst digitaal doorzoekbaar worden gemaakt. In TRIADO zijn tests gedaan met twee bekende OCR (Optical Character Recognition)-pakketten: Abbyy Finereader en Tesseract. Beide haalden de beste resultaten bij volledig getypte documenten zoals processen-verbaal en besluiten. Abbyy Finereader scoorde het best voor dit soort documenten: er werd een 'word error rate' (WER oftewel het percentage foutief omgezette woorden) gemeten van 15.62. Ter vergelijking: voor een kleine steekproef van historische kranten werd door de Koninklijke Bibliotheek een WER van 9 gemeten. Het mindere resultaat voor archiefmateriaal versus drukwerk valt grotendeels te verklaren door de conditie van het papier (doorslagen), gebrekkige typemachines en vervaagde inkt.

Voor een archief zoals het CABR, met een hoge informatiedichtheid en veel feitelijke vermeldingen van namen, plaatsen, datums en gebeurtenissen, is een WER van 15 hoog maar goed genoeg om een zoekfunctie omheen te bouwen en nadere verrijkingen aan te brengen. In TRIADO is vanwege de privacygevoelige informatie in het CABR binnen een streng beveiligde omgeving gewerkt. Het was niet mogelijk om te experimenteren met software die wordt ingezet voor handschriftherkenning. In projecten zoals READ (read.transkribus.eu) worden hiermee inmiddels goede resultaten behaald.

'In TRIADO zijn tests gedaan met twee bekende OCR-pakketten: Abbyy Finereader en Tesseract'

'Autoclassificatie heeft grote potentie om omvangrijke archieven digitaal nader toegankelijk te maken'

Verschillen

Tests in TRIADO wezen uit dat voorbewerking van de scans door het zwarter maken van pixels die op inkt lijken ('darkening'), in combinatie met *machine learning* de kwaliteit van de machineleesbare tekst nog verder kan verbeteren. Met enige handmatige training leerde de computer zichzelf om te leren. Er werden meer fouten gemaakt, maar tegelijkertijd herkende de software ook meer woorden.

Interessant was het verschil tussen Tesseract en Abbyy. Tesseract heeft de neiging om van alles op een pagina letters en woorden te maken. Abbyy daarentegen laat twijfelgevallen buiten beschouwing en is erg goed in het omzetten van reguliere getypte tekst. Doordat Tesseract er heel veel extra letters en woorden bij verzint, scoort het lager in de WER-scores. Uiteindelijk herkent het soms wel woorden die Abbyy negeert. In technische termen: de 'precision' (percentage van de gevonden woorden dat correct is) van Abbyy is hoger, maar de 'recall' (percentage woorden dat kan worden teruggevonden) is lager. In TRIADO is ervoor gekozen met verschillende OCR-lagen (Abbyy, Tesseract, Tesseract darkened) te werken om het beste van alle drie methodieken te benutten.

Zoekfilter

Gebruikers zoeken doorgaans op namen van personen, organisaties, gebeurtenissen, locaties en datums. Door deze elementen in de machineleesbare tekst te detecteren, kunnen er zoekfilters worden ontwikkeld. Een experiment in TRIADO om met 'named entity'-software Frog-NER alle woorden met een naam uit de tekst te halen, leverde matige resultaten op: oude Nederlandse spelling en ook Duitse woorden in de tekst leidden tot hoge foutenmarges. Het automatisch herkennen van datums bleek wel goed mogelijk.

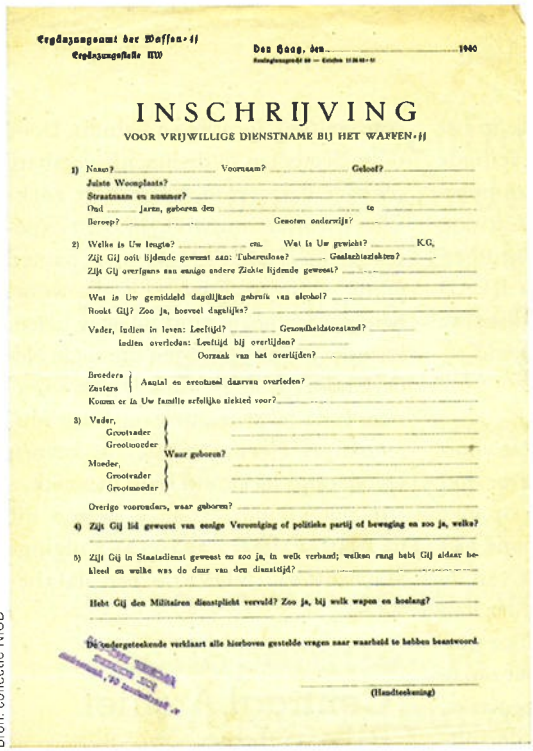
Uit een kleine test met namen uit de Nationale Database Vervolgings Slachtoffers (NDVS) – een databe-

Centraal Archief Bijzondere Rechtspleging

Het Centraal Archief Bijzondere Rechtspleging (CABR) is het meest geraadpleegde Tweede Wereldoorlogarchief in Nederland. Het omvat de dossiers van circa 300.000 personen die verdacht werden van collaboratie met de Duitse bezetter. Dit justitiële archief, beheerd door het Nationaal Archief, bestrijkt circa vier strekkende kilometer aan processen-verbaal, dagvaardingen, inlichtingenformulieren, besluiten, cassatieverzoeken, lidmaatschapskaarten van collaborerende instellingen en allerlei onderliggende bewijsstukken zoals foto's en brieven. Veel is getypt, maar het CABR bevat ook handgeschreven stukken en formulieren waarin beide worden gecombineerd. Op dit moment is het archief beperkt openbaar en alleen ontsloten op de naam van de verdachte. Voor meer informatie: tinyurl.com/yxs2k3ee; over de ontstaansgeschiedenis van het CABR: tinyurl.com/ycumpjo7.

Slotconferentie over TRIADO

Op 13 september organiseert het Netwerk Oorlogsbronnen een slotconferentie over TRIADO. Meld u op oorlogsbronnen.nl aan voor de nieuwsbrief en blijf zo op de hoogte.

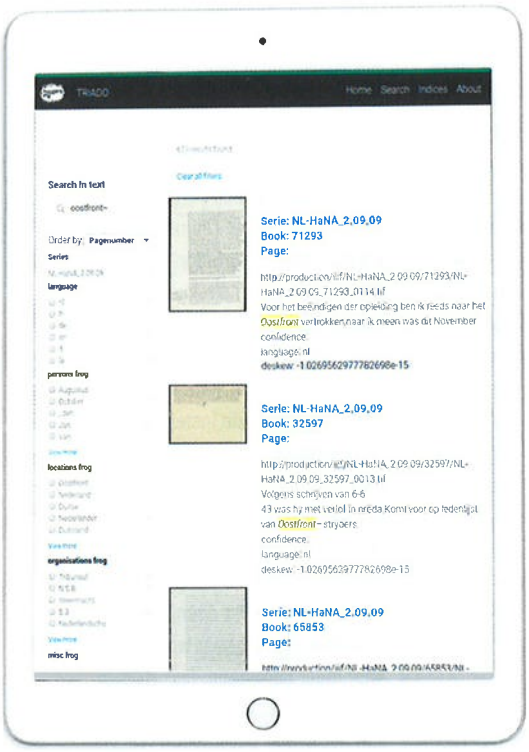


INSCHRIJVING

VOOR VRIJWILLIGE DIENSTNAME BIJ HET WAFFEN-SS

Bron: collectie NIOD

Leeg inschrijfformulier van de Waffen SS



Screenshot van de demonstrator

stand van het Herinneringscentrum Kamp Westerbork en het Joods Cultureel Kwartier met de namen van circa honderdduizend personen – konden in de bijna veertien meter testmateriaal middels een simpele zoekactie al enkele slachtoffers worden geïdentificeerd. De verwachting is dat het matchen van de machineleesbare tekst aan bestaande databestanden zoals bijvoorbeeld de WO2-thesaurus, GeoNames en Wikidata, de digitale toegang tot het materiaal en de verbindingen naar andere informatiebronnen aanzienlijk kan verbeteren.

Autoclassificatie

Een van de belangrijkste onderdelen van TRIADO was de test met het automatisch herkennen van soorten documenten (autoclassificatie). Voor 28 type documenten zijn handmatig voorbeelden verzameld. Vervolgens is de computer getraind om op basis van deze voorbeelden soortgelijke documenten in de rest van de dertien meter te herkennen.

De software heeft zichzelf verder getraind met de gevonden resultaten (*deep learning*). In eerste instantie werd zeventig procent van de documenten correct herkend. Machine learning leverde een verbetering op van circa tien procent. De computer vertoonde opvallend menselijke trekjes: processen-verbaal en getypte correspondentie werden vaak door elkaar gehaald, gedrukte lidmaatschapskaarten konden daarentegen goed worden onderscheiden. Conclusie: autoclassificatie heeft grote potentie om omvangrijke archieven digitaal nader toegankelijk te maken.

Topic modelling

Topic modelling-software analyseert machineleesbare tekst uit documenten en vat het samen in enkele veelvoorkomende steekwoorden. Een van de bijkomstige voordelen van topic modelling is dat door de OCR foutief omgezette woorden er eenvoudig kunnen worden uitgefilterd omdat ze vaak niet meer dan eenmaal voorkomen. Wat overblijft is bijvoorbeeld het volgende: *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>*

Op basis van deze staccato karakterisering (de namen zijn weggelaten) kan je als gebruiker al een goede indruk krijgen waar de hieraan gekoppelde documenten over gaan. In dit geval een proces-verbaal over een voorval waarbij enkele Groningse Landwachters – hulppolitiemannen gerekruteerd vanuit de Nationaal Socialistische Beweging – verdacht worden van hulp bij het oppakken van onderduikers.

Digitaal ontsluiten van archieven

Na afloop van alle experimenten is er een demonstrator gebouwd. Wie nu zoekt op 'Oostfront', krijgt in een paar seconden 619 treffers met verwijzingen naar documenten waar dit woord gebruikt wordt. Dit is een wereld van verschil met de huidige situatie waarin het CABR alleen op dossierniveau is ontsloten op naam van verdachte.

Het TRIADO-project is nog niet afgelopen. Tot juli zal NIOD-onderzoeker Ismee Tames verkennen welke toegevoegde waarde het digitaal ontsluiten van het CABR voor de wetenschap heeft. Nu al, met alle onvolkomenheden, kan geconcludeerd worden dat de inzet van digitale technologie een immense stap voorwaarts betekent bij het toegankelijk maken van archieven. Niet alleen ligt voor het eerst toegang tot op paginaniveau binnen bereik, ook biedt machineleesbare tekst uitstekende aanknopingspunten om de informatie in verschillende archieven aan elkaar te koppelen. Zover zijn we nog niet. Eerst moet er meer gedigitaliseerd worden en vooral machineleesbaar gemaakt. <

Lees verder

> Onderzoeksrapport TRIADO, zie www.oorlogsbronnen.nl/rapport-triado-verrijkingsfase
> Projectpagina: tinyurl.com/yycqpj8m