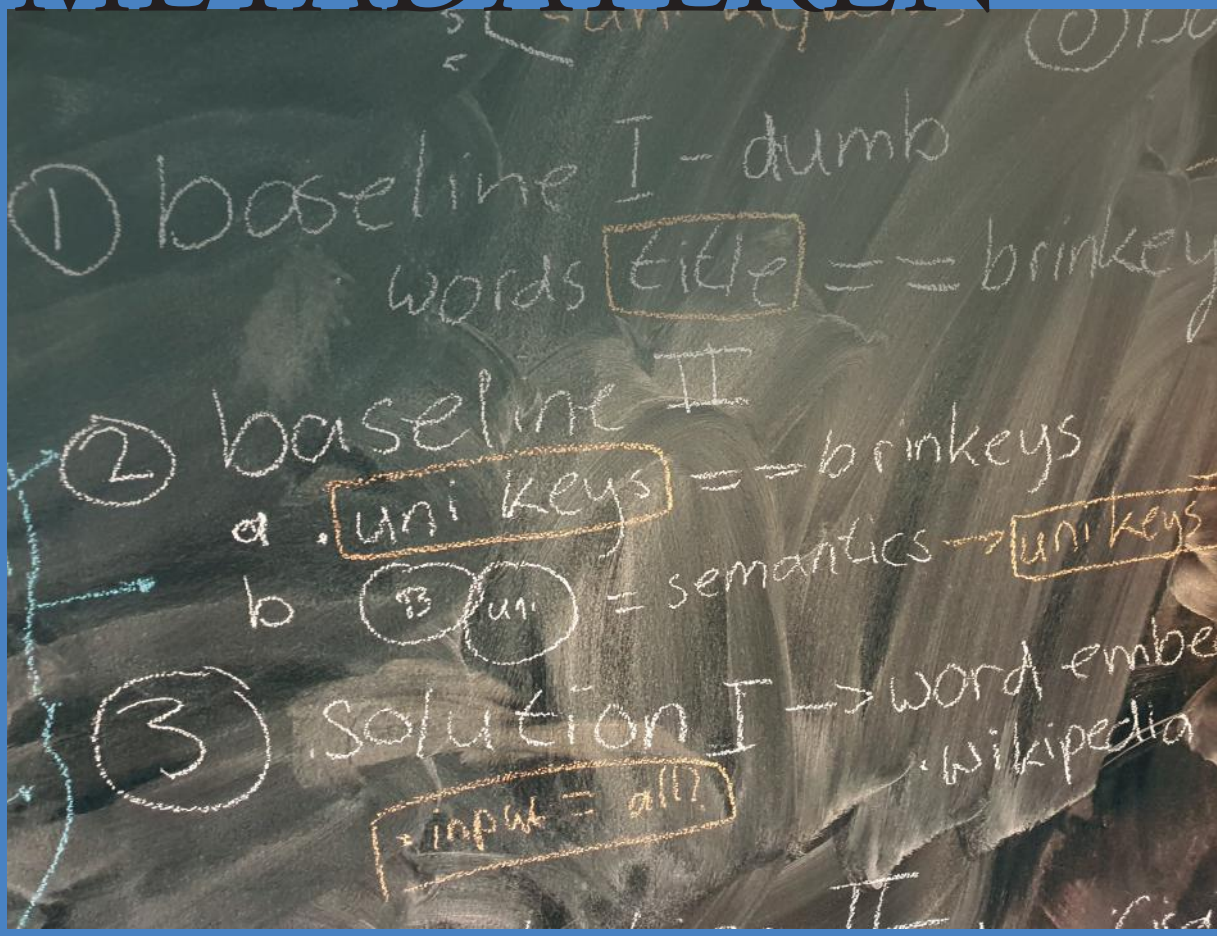
Verkenning mogelijkheden

AUTOMATISCH METADATEREN



Augustus 2019

KB nationale bibliotheek

Colofon

Auteurs: Martijn Kleppe (Orcid 0000-0001-7697-5726; ISNI 0000 0003 8995 1058), Sara Veldhoen (Orcid 0000-0002-8376-0886; ISNI 0000 0004 7703 0857), Meta van der Waal-Gentenaar, (ISNI 0000 0003 9600 2854), Brigitte den Oudsten (ISNI 0000 0004 7703 0865), Dorien Haagsma (ISNI 0000 0004 7703 0873), allen in dienst van de KB.

Eindredactie: Erik Jan Harmens (ISNI 0000 0000 3808 057X)

Vormgeving: Carlien Keilholtz

DOI: http://doi.org/10.5281/zenodo.3373316

Foto omslag: Martijn Kleppe

Jaar van uitgave: 2019

Plaats van uitgave: Den Haag

INHOUDSOPGAVE

INLEIDING

Hoe kunnen we het beschrijven van publicaties vergemakkelijken met behulp van slimme technieken? Dit is een van de onderzoeksvragen van de KB Onderzoeksagenda waar we de komende jaren een antwoord op willen geven[1].

Momenteel vindt het beschrijven van publicaties, ook wel *metadateren* of *titelbeschrijven* genoemd, binnen de Koninklijke Bibliotheek (KB) deels handmatig plaats en deels door het overnemen van gegevens die we verkrijgen via andere bronnen. Mede door de groei van digitaal vervaardigd materiaal (*born digital*) en de groeiende opslag van websites, verwachten we de komende jaren meer publicaties dan voorheen te willen bewaren. Daarom verkennen we de mogelijkheden om het handmatig beschrijven van publicaties te optimaliseren. Twee ontwikkelingen bieden kansen: de groeiende stroom volledig digitaal beschikbare publicaties en het feit dat we door slimme technieken uit onder andere de Kunstmatige Intelligentie, waaronder *machine learning*, steeds beter in staat zijn om die digitale teksten door de computer te laten interpreteren.

In dit whitepaper geven we de stand van zaken van onze eerste verkenningen van de mogelijkheden van het automatisch metadateren van publicaties. Eerst geven we een overzicht van de manieren waarop organisaties en bedrijven buiten de KB bronnen zoals nieuwsartikelen, boeken, tv-uitzendingen of foto's slim analyseren en beschrijven. Daarna bespreken we hoe we op dit moment binnen de KB titels beschrijven, om aan te kunnen geven waar in het proces we de mogelijkheden van automatische metadatering verkennen. In het derde hoofdstuk bespreken we resultaten van onze eigen experimenten met het automatisch toekennen van trefwoorden aan publicaties. We sluiten af met de lessen die we tot nu toe geleerd hebben en beschrijven onze volgende stappen.

Onze missie is om vanuit de kracht van het geschreven woord als netwerkorganisatie bij te dragen aan een slimmer, vaardiger en creatiever Nederland. Daarom delen we onze bevindingen niet alleen binnen de KB, maar ook daarbuiten, met iedereen die met soortgelijke ontwikkelingen bezig is of daarin interesse heeft. Alleen door kennis te delen en samen te werken kunnen we iedereen in staat stellen om te lezen, te leren en onderzoek te doen.

Metadateren

het beschrijven van publicaties. Het wordt ook wel *titelbeschrijven* genoemd.

AUTOMATISCH METADATEREN

elders

Het automatisch analyseren van de inhoud van media-uitingen gebeurt al langere tijd. Op het gebied van kunstmatige intelligentie en machine learning zijn op dit moment de spraakassistenten in bijvoorbeeld mobiele telefoons wellicht het meest bekend. Op basis van een gesproken vraag zijn smartphones in staat om spraak om te zetten in tekst en die vervolgens te analyseren op veelgebruikte termen, personen, locaties of concepten.

Voorheen werden dit soort toepassingen vooral gemaakt door de grote techbedrijven, maar door de beschikbaarheid van steeds meer grote datacollecties, open source software en rekenkracht zien we ook toepassingen van dit soort technieken in andere domeinen. In dit hoofdstuk laten we een aantal van deze toepassingen in voor de KB aanpalende sectoren de revue passeren en beschrijven we wat we zelf op dit gebied al hebben gedaan.

Media

Om ze beter vindbaar te maken, laten journalisten hun artikelen en foto's op verschillende manieren automatisch analyseren door de computer. Het Amerikaanse tijdschrift Forbes werkt met een content-managementsysteem dat op basis van de inhoud van een artikel trefwoorden suggereert aan een journalist terwijl hij of zij een artikel schrijft.[2] Getty past soortgelijke technieken toe om redacteuren die werken aan een verhaal te helpen met het vinden van een foto bij dat verhaal.[3] De New York Times werkt samen met Google Cloud om hun fotoarchief te ontsluiten, waarbij foto's behalve gedigitaliseerd ook automatisch worden beschreven.[4] In Nederland werkt RTL aan toepassingen om spraak- en beeldherkenning te gebruiken voor het doorzoeken van tv-content.[5]

Naast het vergemakkelijken van het productieproces en het zoeken naar redactionele inhoud werken verschillende mediapartijen ook aan toepassingen voor consumenten. Al lange tijd wordt er gewerkt aan zogeheten aanbevelers (recommenders), die op basis van de inhoud van programma's of nieuwsberichten, soortgelijke of andersoortige artikelen, series of muziek aanbevelen. Bekende voorbeelden zijn de aanbevelingen die Spotify en Netflix doen op basis van eerder beluisterde muziek of bekeken series en films. Er zijn ook recommenders die aanbevelingen doen op basis van tekstuele inhoud. In Nederland werkt Het Financieele Dagblad hiermee[6] , Blendle doet het in hun gepersonaliseerde nieuwsbrief.[7] Daarnaast werkt de Nederlandse Publieke Omroep (NPO) aan een recommender die televisieprogramma's aanbeveelt op basis van de publieke waarde van een programma.[8]

Dergelijke toepassingen zien we ook bij uitgevers van boeken, in Nederland bijvoorbeeld bij Bookarang. Waar aanbevelers bij webshops zich beperken tot statistieken gebaseerd op eenvoudige transacties ("Mensen, die dit boek kochten, kochten ook deze boeken"), is Bookarang in staat om aanbevelingen te doen op basis van de inhoud van boeken. Hun techniek wordt onder andere gebruikt door de AKO en in de Online Bibliotheek.[9] WPG Uitgevers heeft samen met Driven by Data de soortgelijke toepassing Thrill Seeker gemaakt, specifiek voor lezers die aanbevelingen willen op het gebied van thrillers.[10] Kenmerkend aan deze recommender is de geboden transparantie: de lezer krijgt inzicht in waarom en op welke onderdelen een aanbeveling wordt gedaan. Dat kan zijn het subgenre, de inhoud van het verhaal, het emotieverloop, de aanwezigheid van bepaalde trefwoorden of het thema. Tenslotte is Google's Talk to Books geen recommender, maar maakt het wel gebruik van soortgelijke technieken en stelt het de gebruiker in staat om een vraag te formuleren, waarna de machine learning-technieken van Google op zoek gaan naar een citaat uit een boek dat zich bevindt in de index van Google Books.[11]

Erfgoed

Binnen de erfgoedsector worden momenteel interessante stappen gezet om gedigitaliseerd en digitaal geboren erfgoed automatisch te analyseren. Het Nationaal Archief heeft een eerste experiment gedaan met het classificeren van binnenkomende e-mails die het bewaart.[12] Ook werkt het Nationaal Archief samen met het Huygens Instituut voor Nederlandse Geschiedenis en het Netwerk Oorlogsbronnen binnen het onderzoeksproject TRIADO, onder andere aan het automatisch classificeren van documenten uit het Centraal Archief Bijzondere Rechtspleging (CABR). De onderwerpen van die

documenten worden met behulp van deep learning-technieken vastgesteld.[13] Naturalis Biodiversity Center past dezelfde soort machine learning-technieken toe voor het automatisch classificeren van afbeeldingen van insecten.[14] Hiervoor werken zij samen met onder andere Waarneming.nl en Observation.org, waar vrijwilligers hun afbeeldingen met beschrijvingen van insecten uploaden. Deze data worden gebruikt als trainingsmateriaal om een bepaald algoritme te ontwikkelen, genaamd classifier, dat op basis van de beschrijvingen van vrijwilligers zelfstandig insecten gaat herkennen.[15] Ook bij de KB werken we samen met (wetenschappelijke) partners aan soortgelijke onderzoeken, zowel binnen onderzoeksprojecten als in het kader van ons researcher-in-residence-programma, waarvan we de resultaten delen in ons KB Lab: https://lab.kb.nl/.[16] Zo hebben we samen met Frank Harbers van de Rijksuniversiteit Groningen de mogelijkheden verkend van het automatisch toekennen van journalistieke genres aan historische krantenartikelen en werken hier verder aan in een gezamenlijk project met onder andere het Centrum Wiskunde & Informatica (CWI), CLARIAH en het eScience Center, genaamd NEWSGAC.[17] Met Thomas Smits (Universiteit Utrecht) en Melvin Wevers (KNAW Humanities Cluster) verkenden we de mogelijkheden van beeldherkenning om bijvoorbeeld afbeeldingen in historische kranten te kunnen onderverdelen in foto's, tekeningen en cartoons, maar ook om de afbeelding te analyseren.[18] Met Puck Wildschut (Radboud Universiteit) experimenteerden we met het herkennen van personages in romans en het in kaart brengen van hun relaties.[19]

De instelling in Nederland die op dit gebied misschien wel het meest ver is, is het Nederlands Instituut voor Beeld en Geluid. Vanwege hun jarenlange ervaring in nationale en Europese onderzoeksprojecten, zijn zij sinds 2018 in staat om voorbij het experiment te gaan en het automatisch beschrijven van tv- en radioprogramma's te implementeren in hun productieproces.[20] Zo wordt spraakherkenningssoftware gebruikt om tv- en radioprogramma's om te zetten in tekst, automatic speaker labeling om namen toe te kennen aan de gegenereerde teksten en gezichtsherkenning om personen in programma's te identificeren.[21] Ten slotte worden op basis van de gegenereerde teksten onderwerpen toegekend aan de programma's op basis van een trefwoordensysteem (thesaurus) dat Beeld en Geluid gebruikt.[22] Deze laatste toepassing lijkt het meest aan te sluiten bij ontwikkelingen binnen het bibliotheekdomein.

Bibliotheken

Al sinds het beschikbaar komen van digitale teksten in de jaren vijftig, worden binnen de bibliotheekwereld de mogelijkheden verkend van automatische titelbeschrijving om het beschrijfproces van publicaties te vergemakkelijken. Wie artikelen leest van Robert David Stevens (University of Manchester) of Karen Spärk Jones (Cambridge University Computer Laboratory) krijgt een mooi overzicht van de ontwikkelingen uit die beginperiode. In dit whitepaper beperken we ons tot de meest recente ontwikkelingen. Goede bronnen zijn voor ons de International Federation of Library Associations and Institutions (IFLA), de Ligue des Bibliothèques Européennes de Recherche (LIBER), het Dublin Core Metadata Initiative (DCMI) en Semantic Web in Libraries (SWIB).

We zien op het gebied van het vergemakkelijken van het beschrijfproces van publicaties twee ontwikkelingen: allereerst worden de mogelijkheden verkend om trefwoorden uit de eigen thesaurus automatisch toe te kennen. De Deutsche Nationalbibliothek gaf tijdens de IFLA-conferentie van 2018 een mooi overzicht van hun vorderingen op dit gebied. Momenteel maken zij gebruik van commerciële software die met behulp van een support vector machine algoritme patronen in teksten opspoort en op basis daarvan trefwoorden toekent.[23] De Deutsche Nationalbibliothek is zich bewust van mogelijke fouten en de wens van gebruikers om volledig correcte informatie te krijgen, daarom geven zij de automatisch gegenereerde trefwoorden in hun catalogus weer in een apart veld met de toevoeging: "maschinell ermittelt".

Ook in Nederland zijn er plannen voor het toekennen van trefwoorden en andere metadata met behulp van commerciële software. Het eerder genoemde Bookarang werkt bijvoorbeeld samen met NBD Biblion om de mogelijkheden hiervan te verkennen.[24] Zij werken aan een tool die op basis van een pdf van een boek een titelbeschrijving genereert, inclusief trefwoorden en een recensie.

De Zweedse Nationale Bibliotheek maakt gebruik van een soortgelijke aanpak, maar dan met eigen ontwikkelde software. Een eerste inzicht was dat de dataset om materiaal te trainen te beperkt in omvang was, een punt dat Joseph Busch ook maakt in zijn paper "Categorization Ethics: Questions about Truth, Privacy and Big Data".[26] De Noorse Nationale Bibliotheek heeft dit geprobeerd op te lossen door een aantal documenten automatisch te genereren.[27] Daarnaast lopen de Noren voorop met inspanningen om multimediale collecties slim te analyseren, zowel voor klanten als voor interne processen.[28]

De Finse Nationale Bibliotheek heeft veel energie gestoken in het creëren van Annif: een open-source-systeem om trefwoorden te genereren op basis van een eigen thesaurus en trainingsdata.[29] Annif maakt gebruikt van een aantal open source-tools voor Natural Language Processing en machine learning-pakketten als Maui, FastText en Gensim. Het wordt actief ontwikkeld en kent een groeiende gebruikersgroep.[30]

Naast het automatisch toekennen van trefwoorden uit de eigen thesaurus, zien we een tweede ontwikkeling: verschillende bibliotheken verkennen de mogelijkheden om extra metadata toe te voegen om het voor de gebruikers nog eenvoudiger te maken om publicaties te vinden. Zo werkt de National Library Board van Singapore sinds 2016 aan het herkennen en duiden van namen van personen uit teksten met behulp van *Named Entity Recognition* (NER). In 2017 worden ook onderwerpen gegenereerd uit teksten met behulp van gecontroleerde databases zoals Wikidata en Geonames, waardoor verschillende collecties nu doorzocht kunnen worden met behulp van een geïntegreerde interface.[31] De George A. Smathers Library van de Universiteit van Florida koos een soortgelijke aanpak, maar dan toegepast op bijzondere collecties. Metadatavelden werden uitgebreid met de thesaurus van de digitale JSTOR-collectie en ook geografische locatiegegevens, waardoor de doorzoekbaarheid en vindbaarheid toenamen.[32] Binnen de KB is hier ook veel mee geëxperimenteerd. Zo hebben Theo van Veen, Juliette Lonij en Willem Jan Faber de afgelopen jaren hard gewerkt aan het herkennen van entiteiten, zoals namen en plaatsen, in historische kranten met behulp van Named Entity Software. Deze entiteiten zijn vervolgens gekoppeld aan databases als DBpedia en Wikidata. De extra informatie die dit opleverde is in een interne en experimentele omgeving van Delpher.nl ondergebracht. Gebruikers kunnen nu zoeken op kenmerken van een persoon, zonder dat die kenmerken genoemd worden in de krant.[33] Een soortgelijke aanpak zien we ook in de tool Text Analyzer van JSTOR.[34] Hierin kunnen gebruikers een tekst laden, waarna het systeem de tekst analyseert, relevante entiteiten eruit haalt en op basis daarvan suggesties doet van relevante artikelen in de database van JSTOR.

Metadata genereren en MOGELIJKHEDEN TOT AUTOMATISEREN binnen de KB

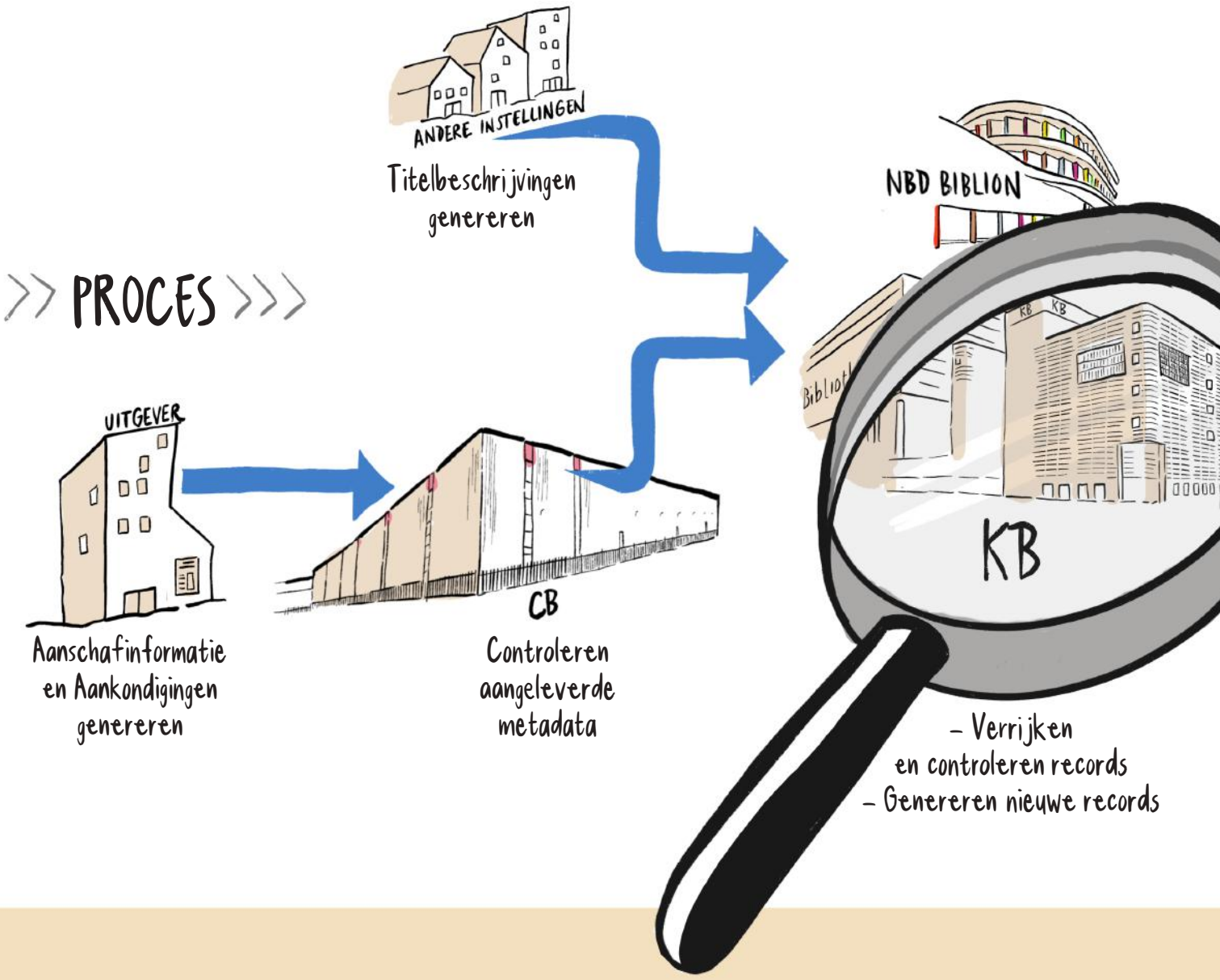
Na een inventarisatie van de mogelijkheden van automatisch metadateren, beschrijven we in dit hoofdstuk hoe we momenteel binnen de KB de publicaties beschrijven en van metadata voorzien. We zullen aangeven waar in het proces we de mogelijkheden verkennen om metadata slimmer te produceren.

Korte beschrijving proces

Binnen de KB genereren we metadata op verschillende manieren. Een deel van de processen vindt extern plaats (zie sterk vereenvoudigde weergave op de volgende bladzijden). Van alle publicaties die bij de KB binnenkomen heeft ongeveer 70% al een record in het Gemeenschappelijk Geautomatiseerd Catalogiseersysteem (GGC). Het grootste deel van deze records zijn metadata, afkomstig van uitgevers en geleverd via het Centraal Boekhuis : de zogenaamde ONIX-metadata.[36] Een heel klein aantal records is nog afkomstig van andere organisaties, tot enkele jaren geleden catalogiseerden bijvoorbeeld ook universiteitsbibliotheken nog in het GGC.

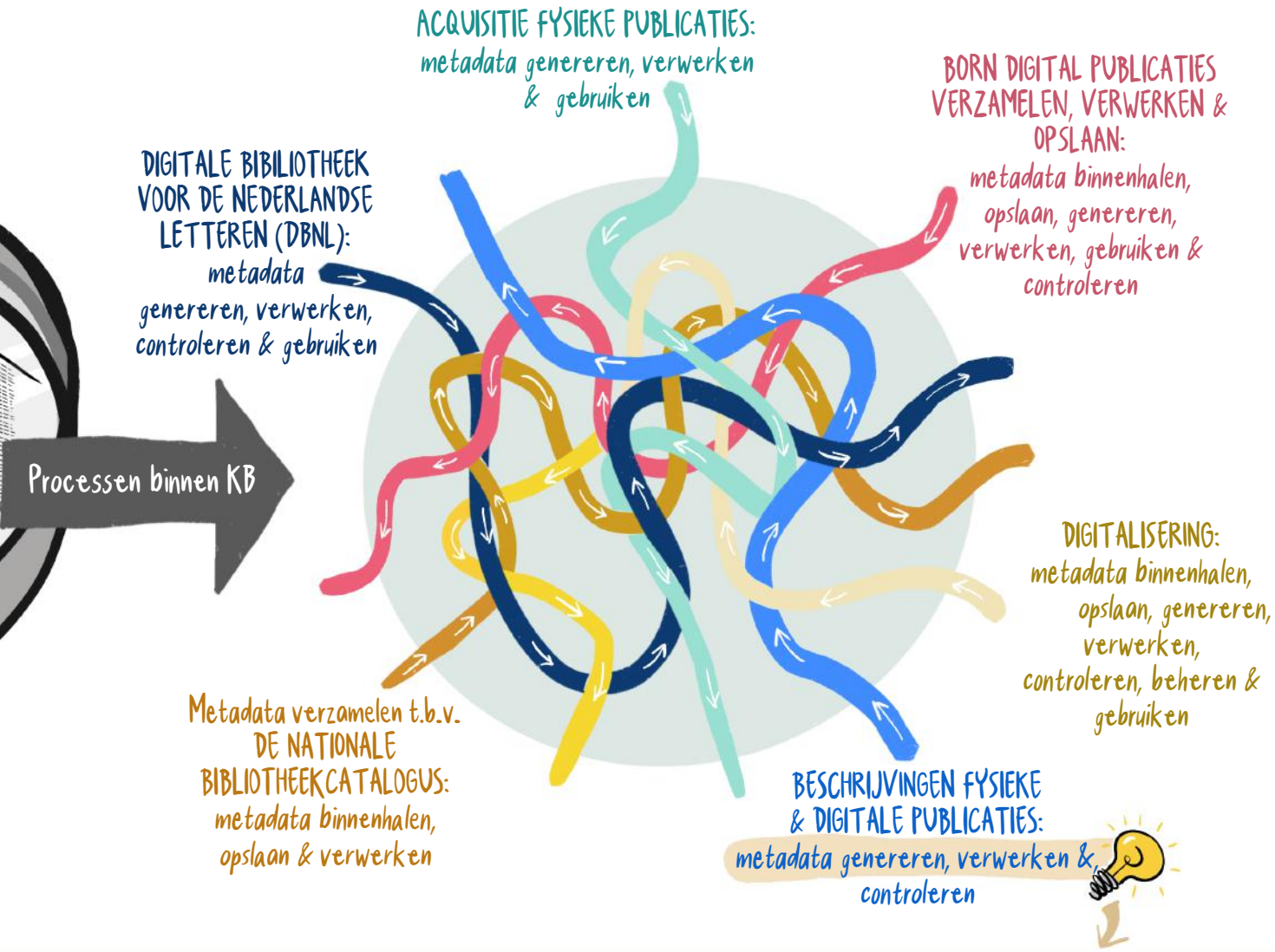
De metadata in de records geleverd via het Centraal Boekhuis zijn niet compleet. De gegevens zijn bijvoorbeeld nog niet gekoppeld aan een thesaurus en soms moeten meer gegevens over de inhoud toegevoegd worden. NBD Biblion[37] en de KB verrijken en controleren de metadata. Wanneer nog geen metadata aanwezig zijn, genereren de titelbeschrijvers van de KB zelf een record. Naast het genereren van metadata tijdens het catalogiseren, worden ook in andere processen nieuwe metadata gegenereerd.

VERKENNING MOGELIJKHEDEN



AUTOMATISCH METADATEREN

binnen de Koninklijke Bibliotheek



UITGEVOERDE VERKENNING:

Innovatie bij Brinkman-trefwoorden
(bij genereren en verwerken metadata)

TOEKOMSTIGE VERKENNINGEN:

nader te bepalen ...

Kwaliteit metadata

Voor de KB is de kwaliteit van de metadata belangrijk, omdat we duurzaam opslaan belangrijk vinden en omdat we als nationale bibliotheek ervoor willen zorgen dat er ten minste voor alle Nederlandse publicaties betrouwbare informatie over auteurs en titels beschikbaar is: de Nationale Bibliografie. De Nationale Bibliografie geldt als het gaat om Nederlandse titels als hét referentiebestand in de nationale en internationale bibliotheeksector en binnen de wetenschappelijke wereld.

Beschrijven of metadateren heeft als doel dat gebruikers informatie goed kunnen vinden, identificeren, selecteren, verkrijgen en gebruiken.[38] Metadata zijn onder andere *identifiers*, titelgegevens, informatie over auteur, publicatie-type en informatiedrager, gegevens over de metadata zelf en administratieve gegevens zoals vindplaats, annotaties en onderwerpsontsluiting. De KB definieert metadata volgens de internationale standaard Resource Description and Access (RDA).[39]

Beschrijven blijft nodig. Het is mogelijk digitale publicaties *full text* te doorzoeken, maar lang niet alle publicaties zijn al digitaal beschikbaar. Dankzij metadata kan een gebruiker gemakkelijk meerdere publicaties vinden met hetzelfde onderwerp, in meerdere talen of op mogelijk meerdere schrijfwijzen. Een onderwerp kan woordelijk niet in een publicatie vermeld worden, maar er wel op van toepassing zijn. Een trefwoord maakt ook deze publicatie vindbaar voor de gebruiker. Daarnaast kan een gebruiker dankzij metadata op eenvoudige wijze meerdere edities van een publicatie vinden of de meest recente editie achterhalen. Tenslotte voegt het context toe aan een publicatie.[40] Een betrouwbare Nationale Bibliografie kan alleen gezaghebbend zijn met goede en gecontroleerde metadata en een gecontroleerde koppeling met diverse thesauri.

Onderwerpsontsluiting

Onderwerpsontsluiting betreft het toegankelijk maken van de inhoudelijke aspecten van een publicatie met behulp van trefwoorden of classificaties. De meeste meerwaarde heeft onderwerpsontsluiting als gecontroleerde trefwoorden, termen en codes worden gebruikt uit een thesaurus of classificatie, want dan is het mogelijk groeperingen te maken die de gebruiker overzicht geeft over alle voor hem of haar relevante publicaties.

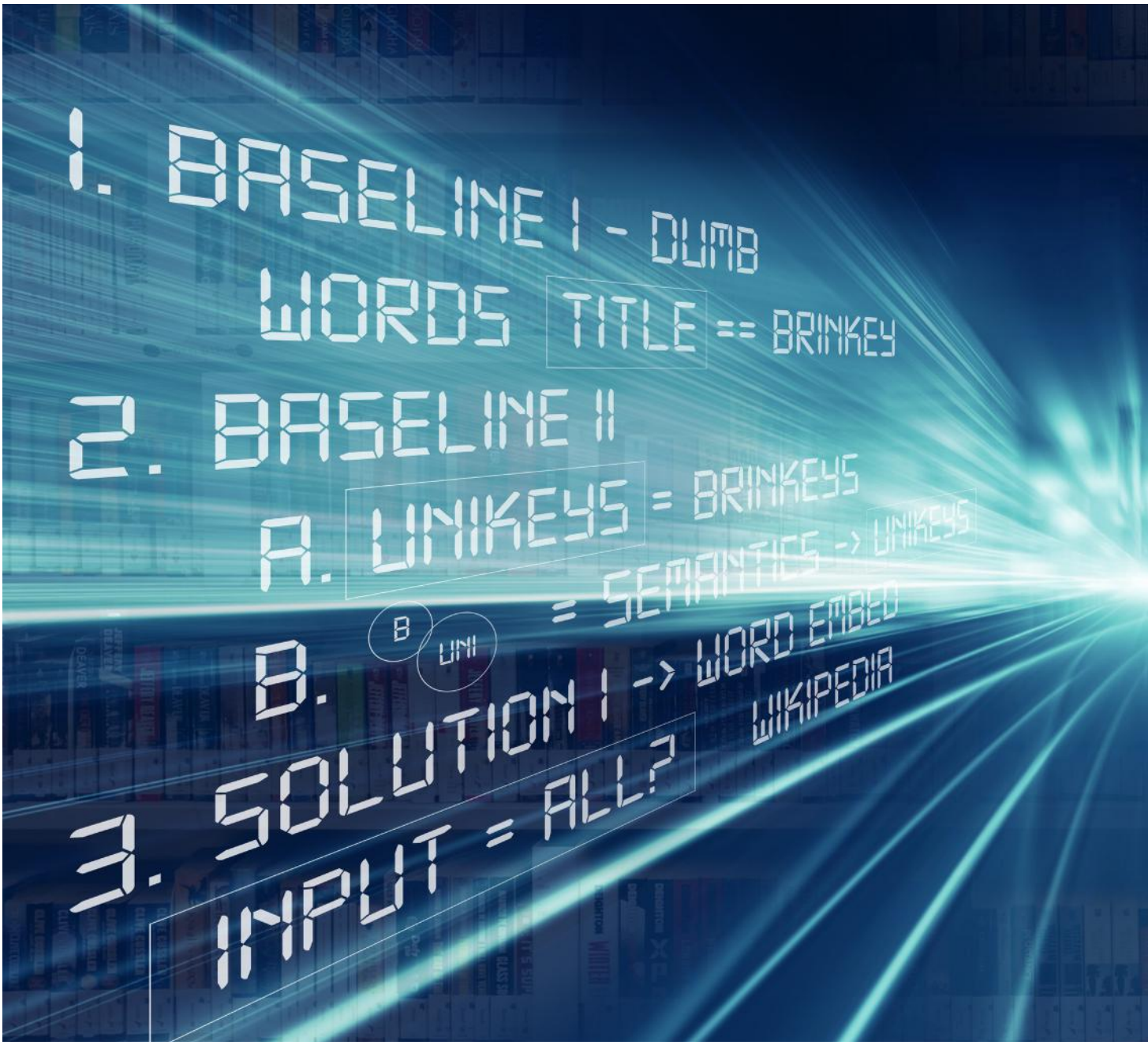
Onderwerpsontsluiting verbetert de kwaliteit van de beschrijvende of bibliografische metadata. Het handmatig toevoegen is een tijdrovende activiteit, daarom ontsluit de KB alleen non-fictie op onderwerp met een eigen Nederlandstalige thesaurus: de Brinkman-onderwerpen. Dit is een thesaurus die gebruikt wordt voor de onderwerpsontsluiting voor een belangrijk deel van de collectie van de KB. Aan kinderboeken worden ook specifieke kenmerken en genres uit een thesaurus toegevoegd, ten behoeve van het Centraal Bestand Kinderboeken. Fictie krijgt echter alleen een trefwoord voor genre of vorm toegekend uit de Brinkman-onderwerpen. In het geval er wel sprake is van onderwerpsontsluiting voor fictie, is die hoofdzakelijk afkomstig van NBD Biblion.

Mogelijkheden automatisch genereren metadata

In het onderzoek naar mogelijkheden voor het automatisch genereren van metadata is gekeken naar waar het meest behoefte aan zou zijn en wat voor de KB de meeste efficiency oplevert. Verbeteren van de aangeleverde metadata door met externe organisaties in de keten te overleggen, viel buiten de scope van dit onderzoek. Wel is de KB met partijen in gesprek over initiatieven die betrekking hebben op het automatisch genereren van metadata, zoals in het vorige hoofdstuk beschreven. Dit kan betere en completere metadata opleveren, al zullen we altijd een zekere mate van redactie willen uitoefenen om de kwaliteit van onze metadata te garanderen.

Bestaande processen waarbij metadata al automatisch worden gegenereerd, vielen ook buiten de scope van dit onderzoek. Het gaat dan bijvoorbeeld om de verwerkingsprocessen van born digital-publicaties en gedigitaliseerd materiaal, waarbij onder andere technische metadata (bijvoorbeeld een bestandsformaat of controlegetal) en administratieve metadata worden gegenereerd.

We hebben gekeken naar welke metadata geautomatiseerd gegenereerd zouden kunnen worden en welke niet, zoals koppelingen met een thesaurus, annotaties en administratieve gegevens. De resultaten van dat onderzoek beschrijven we in het volgende hoofdstuk.



Resultaten:
AUTOMATISCH TOEKENNEN
van trefwoorden

De eerste verkenning die we zelf hebben gedaan, concentreerde zich op onderwerpsontsluiting door middel van het automatisch toekennen van trefwoorden. Hiervoor gebruikten we de Brinkman-onderwerpen. Het toekennen van Brinkman-onderwerpen gebeurt momenteel handmatig. Een titelbeschrijver heeft de publicatie op zijn of haar bureau liggen en maakt een inschatting van de toe te kennen trefwoorden. Omdat we deze publicaties ook steeds vaker volledig digitaal ter beschikking hebben, onderzochten we in hoeverre we de computer konden trainen om automatisch suggesties van dit soort Brinkman-onderwerpen, door het onderzoeksteam ook wel Brinkeys genoemd, te doen.

Deelnemers aan de ICT with Industry Workshop, vlnr: Martijn Kleppe (KB), Rob Koopman (OCLC Research), Karin Goes (VU), Shenghui Wang (OCLC Research), Areumbyeol Kim (VU), Myrthe Reuver (Radboud Universiteit), Iris Hendrickx (Radboud Universiteit), Sara Veldhoen (KB), Alex Brandsen (Universiteit Leiden), Hugo de Vos (Universiteit Leiden), Sepideh Mesbah (TU Delft), Hugo Huurdeman (UvA), Richard Zijdeman (IISG).

ICT With Industry Workshop

De eerste stappen hiertoe hebben we gezet tijdens de ICT with Industry Workshop in januari 2019. Dit was een door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek georganiseerde workshop in het Lorentz Center in Leiden, waar bedrijven en maatschappelijke organisaties samen met een groep wetenschappers een casus uitwerken.[43] Het team onderzoekers dat aan de casus van de KB werkte had verschillende achtergronden en expertises en varieerde qua samenstelling van masterstudenten tot senior onderzoekers. Iris Hendrickx, als onderzoeker verbonden aan de Radboud Universiteit in Nijmegen, was de academisch leider van het project. De resultaten die we beschrijven in dit hoofdstuk zijn gebaseerd op het rapport dat het team na afloop heeft opgesteld.[44]

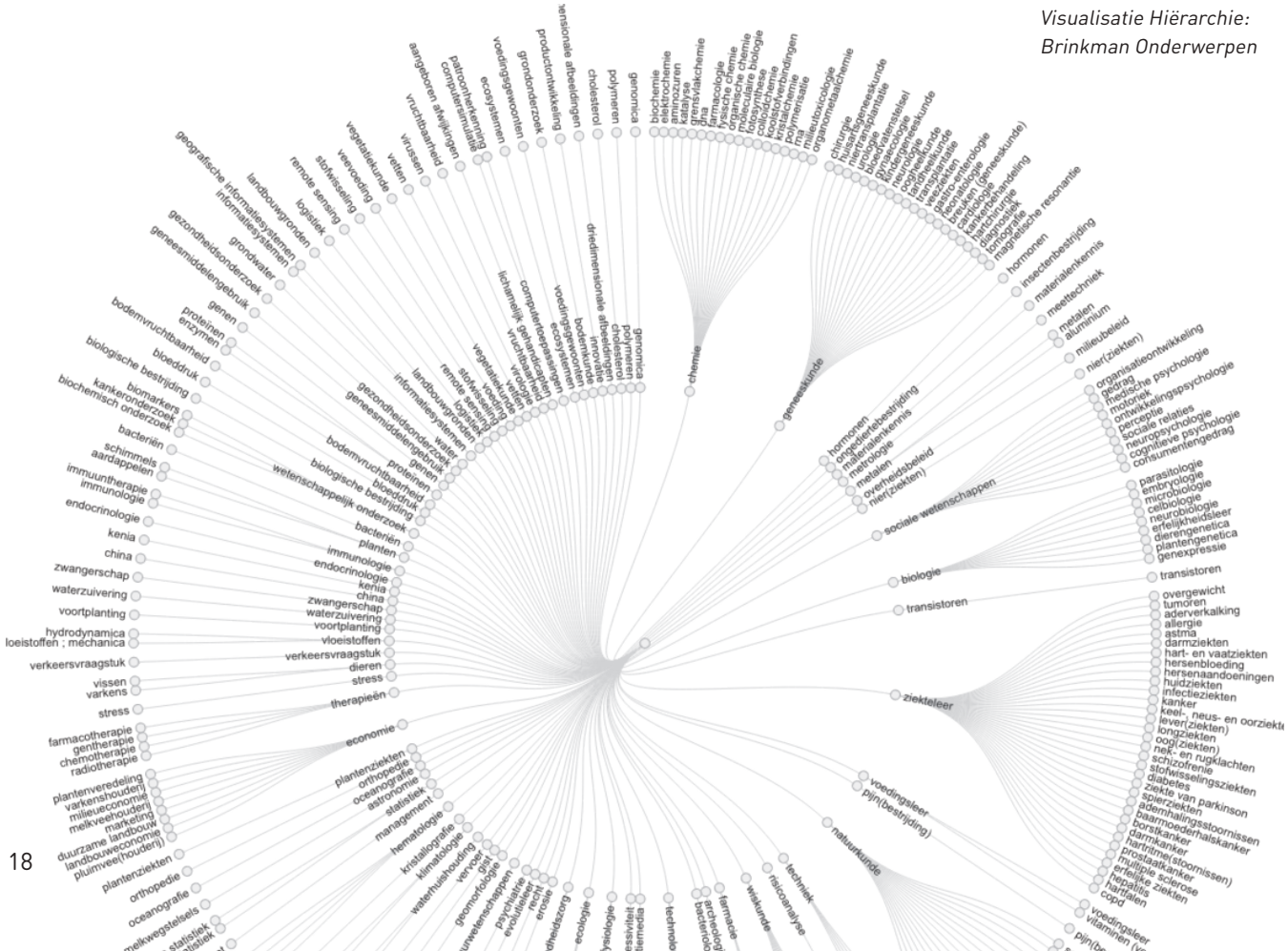
Ons doel voor de week was een verkenning van verschillende methoden voor het automatisch toekennen van trefwoorden aan digitale publicaties. Omdat we werkten met publicaties buiten de muren van de KB, experimenteerden we met dissertaties die als open access-publicatie zijn uitgegeven en zijn opgeslagen in het digitale depot van de KB. Omdat we moeite hadden de publicaties uit het digitale depot te halen, hebben we de proefschriften met toestemming van de universiteits-bibliotheken opnieuw binnengehaald vanuit de digitale archieven van de universiteitsbibliotheken van Groningen, Delft, Rotterdam, Leiden, Utrecht en Wageningen. Vanuit deze bronnen waren ook metadata beschikbaar, deze waren echter sterk wisselend van aard en kwaliteit.

Uitdagingen

Om aan het werk te kunnen met deze data hadden we een paar uitdagingen:

- Het koppelen van de titelbeschrijvingen uit het Gemeenschappelijk Geautomatiseerd Catalogiseersysteem (GGC) aan de dissertaties was geen sinecure, omdat vanuit de gegevensbronnen verschillende identifiers (bijvoorbeeld titels en codes) waren toegekend. Indien bekend gebruikten we het ISBN, in andere gevallen de auteursnaam (achternaam) in combinatie met de titel. Voor de laatste twee gebruikten we een zogenaamde *string distance*, want bij het lexicaal vergelijken kan veel mis gaan met bijvoorbeeld bijzondere letters en accenten.
- De dissertaties zijn deels in het Nederlands, maar voor een groot deel in het Engels geschreven, terwijl de Brinkman-onderwerpen in het Nederlands zijn opgesteld. Om dit probleem op te lossen is een hoop tijd en aandacht besteed aan het vertalen van termen en titelwoorden.
- De Brinkman-onderwerpen zijn in beginsel hiërarchisch ingericht, een bruikbaar gegeven voor het automatiseren van de onderwerpsontsluiting. De diepte van de hiërarchie bleek echter nogal wisselend, zoals weergegeven in onderstaande visualisatie.
- Er is een koppeling gemaakt tussen de Brinkman-onderwerpen en Wikidata, omdat laatstgenoemde meer verbanden tussen onderwerpen legt en tevens meertalig beschikbaar is.

Visualisatie Hiërarchie: Brinkman Onderwerpen



Aanpak

Er werden in deze week verschillende aanpakken geprobeerd. Allereerst de "naïeve" *baselines*, om een idee te krijgen hoe moeilijk dit probleem is, vervolgens onderzochten we een aantal meer geavanceerde methodes en tot slot lieten we twee voor deze toepassing ontwikkelde tools van externe partijen los op de data.

Bij de eerste naïeve baseline keken we naar de l**exicale overlap van titelwoorden** en **Brinkeys**. Wanneer er letterlijke Brinkeys voorkomen in de titel worden die aangemerkt als (mogelijk) onderwerp. Engelse titels moesten hiervoor worden vertaald, aangezien de thesaurus Nederlandstalig is. Het vertalen was geen sinecure om voor elkaar te krijgen, en leverde onverwacht een lagere score op.

Als tweede naïeve baseline keken we naar de **lexicale overlap van 'Unikeys'** (onderwerpen toegekend door universiteiten) en Brinkeys. Wederom is hier gewerkt met vertalingen van Engelse termen, met de bijbehorende problemen en bezwaren. Deze methode scoort lager dan de vorige, wat deels verklaarbaar is doordat de Unikeys niet volgens een gecontroleerd vocabulaire worden toegekend. Bovendien is aannemelijk dat catalogiseerders zich op de titel baseren, wat tot een hoge overlap leidt.

De eerste methode die we onderzochten was **Naive Bayes**, een eenvoudig machine learning-algoritme dat op basis van woorden die voorkomen in de titel en/of de samenvatting een Brinkey voorspelt. In tegenstelling tot de andere methoden wordt in dit geval altijd precies één onderwerp toegekend. Deze methode leverde erg slechte resultaten op en is snel terzijde geschoven.

Word Embeddings was de tweede methode: een vrij recente techniek die op basis van neurale netwerken de betekenis van woorden in een continue ruimte (*vector space*) plaatst. Zo kun je via woorden of teksten zoeken naar andere woorden (in dit geval: Brinkeys) die qua betekenis het meest verwant zijn. Dit werkt met aanpassingen voor meerdere talen tegelijk, waardoor ook Engelse dissertaties aan de Nederlandstalige Brinkeys gekoppeld kunnen worden. Deze methode gaf hoopvolle resultaten en er zijn een hoop ideeën om deze methode aan te passen en verder te onderzoeken.

FastText was de derde methode: die echter niet specifiek voor deze taak ontwikkeld werd, maar meer in het algemeen voor de classificatie van teksten.[45] De input bestond uit titel, samenvatting, Unikeys en de naam van het instituut waar het onderzoek plaatsvond. Deze methode scoorde vrij hoog.

De eerste tool die we onderzochten was **Annif**: een bestaande oplossing, ontwikkeld door de Finse nationale bibliotheek.[46] De methode biedt de mogelijkheid om een eigen thesaurus van trefwoorden te gebruiken.[47] Onder de motorkap zit een combinatie van bestaande modules voor *natural language processing* en machine learning. Deze modules kunnen op verschillende manieren gecombineerd worden, verder kan een eigen thesaurus worden gebruikt als lijst van mogelijke onderwerpen. We hebben deze tool gedraaid op onze dataset met TF-IDF[48] als wegingsfactor en gebruikmakend van *Snowball-stemming*.[49] Als input gaven we de titel en samenvatting (indien beschikbaar) van elke dissertatie, om op basis hiervan een tiental onderwerpen te voorspellen. De scores kwamen hiermee nog niet zo hoog uit, maar dat zou bij het invoeren van meer data en het experimenteren met andere configuraties flink kunnen verbeteren.

We onderzochten ook **Ariadne**, een recent door het Online Computer Library Center (OCLC) ontwikkelde tool voor het doorzoeken en interpreteren van bibliografische gegevens op basis van tekstkenmerken.[50] Deze aanpak behaalde verreweg de hoogste scores, al weten we nog weinig over de precieze methodiek erachter of het materiaal dat gebruikt is om het systeem te trainen.

Resultaten en demo

Om de resultaten van de workshop te demonstreren heeft het team een demo gebouwd die beschikbaar is op het KB Lab: http://lab.kb.nl/tool/brinkeys-tool. Hier kun je een voorbeelddissertatie naar de *analysebox* slepen, waarna getoond wordt welke Brinkman-onderwerpen er volgens het systeem aan toegekend zouden kunnen worden en welke Brinkman-onderwerpen er door de titelbeschrijvers van de KB aan zijn gehangen.

We hebben de systemen ook beoordeeld. Dit wordt doorgaans bepaald op basis van *precision and recall* .[51] In deze casus ligt onze focus op recall: als het systeem een lijstje met twintig mogelijke Brinkeys teruggeeft, zitten de juiste er volgens onze thesaurus dan tussen? We denken dat een systeem met een hoge recall de titelbeschrijvers kan helpen om snel de juiste onderwerpen te vinden. Daarnaast maten we ook precision: stel dat we een systeem helemaal zelf drie onderwerpen laten toekennen, hoe vaak zijn die toekenningen dan juist?

Method	Recall			Precision	
	At1	At10	At20	At1	At3
Baseline 1 (overlap titel - Brinkey)	16.9			30.5	
Baseline 2 (overlap Unikey - Brinkey)	11.6			14	
Methode 1 (Naive Bayes classifier)	3.5			6.5	
Methode 2 (Multi-lingual word embeddings)			24.8		6.6
Methode 3 (FastText classifier)			40.3		16.2
Tool 1 (Annif)		16.7			16.7
Tool 2 (Ariadne)			56,9		29.2

Door de exploratieve opzet van het onderzoek zijn de hierboven weergegeven resultaten niet eenduidig te interpreteren: er zijn veel variaties geweest in de experimenten. Het is wel duidelijk dat Ariadne buitengewoon goed presteert. Dit is dan ook een tool die specifiek ontwikkeld is voor een dergelijke taak en bovendien vooraf getraind is op een grote hoeveelheid wetenschappelijke literatuur. Het nadeel van deze tool is echter dat deze, in tegenstelling tot Annif, niet open-source beschikbaar is, waardoor we weinig grip hebben op de precieze werking. Ook de FastText classifier scoort behoorlijk hoog en kan binnen Annif ook gebruikt worden als alternatieve backend. In ieder geval kunnen we uit deze hoge scores concluderen dat het zeker mogelijk is om op een zinvolle manier onderwerpen automatisch toe te kennen.

Lessen en VOLGENDE STAPPEN

Deze eerste verkenning leverde mooie resultaten en inzichten op. Toch zijn we nog lang niet zo ver dat we een systeem hebben dat in de praktijk toepasbaar is. In dit hoofdstuk beschrijven we de belangrijkste lessen van onze verkenning en de volgende stappen.

Data, data, data

Het belangrijkste bij experimenten als deze zijn de data waarmee je werkt. Dit hebben we op verschillende niveaus ervaren, allereerst bij het beschikbaar maken van de trainingsdata. Deze hebben we, zoals beschreven, binnengehaald via verschillende bronnen, wat verschillende uitdagingen met zich meebracht op het gebied van dataharmonisatie, maar ook bij het creëren van de *testset*. Omdat in Nederland meerdere partijen bezig zijn met soortgelijke analyses op publicaties, kan het voor ons taalgebied de moeite waard zijn om de mogelijkheden te verkennen van de constructie die de Amerikaanse HathiTrust biedt.[52] Zij slaan digitale publicaties niet alleen duurzaam op, maar stellen deze via hun Research Center ook beschikbaar voor allerlei toepassingen op het gebied van tekst en *datamining*.[53]

Daarnaast was de hoeveelheid data vrij beperkt, zeker als je bedenkt dat er ruim 2200 verschillende Brinkeys zijn toegekend, waarvan sommigen maar sporadisch voorkomen. Bovendien is er tijdens de workshop alleen gewerkt met titels en samenvattingen en nog niet met *full text*. Aan de prestaties van Ariadne kunnen we zien dat er zeker ruimte is om het systeem verder te verbeteren.

Ten slotte zijn de resultaten geëvalueerd op basis van overeenkomst met de onderwerpsontsluiting in de catalogus. Wellicht dat een ander systeem andere Brinkeys suggereert die net zo goed, of zelfs beter passen bij de dissertatie. Om dit te analyseren zouden we het liefst een aantal experts naar een aantal van deze lijsten laten kijken, om op die manier het trainingsmateriaal verder te verfijnen.

Vervolg

Op basis van deze lessen werken we momenteel aan een aantal mogelijke vervolgstappen. Allereerst verkennen we de mogelijkheden van Annif met andersoortige data. Het voordeel van Annif is dat het open source is, meerdere technieken combineert en verschillende toepassingen kent. De resultaten met teksten van proefschriften waren hoopgevend, we weten echter nog niet hoe het werkt met andersoortige teksten. Daarom bekijken we nu welke resultaten Annif geeft als we werken met bijvoorbeeld samenvattingen van boeken of volledige teksten. We zijn benieuwd of de kwaliteit van de gesuggereerde Brinkman-onderwerpen verbetert als een volledig boek wordt geanalyseerd, of dat het analyseren van een deel van het boek volstaat.

Een andere richting waar we momenteel naar kijken is het automatisch toevoegen van metadata aan de teksten op de website van de Digitale Bibliotheek voor de Nederlandse Letteren (DBNL).[54] Op deze website bevinden zich teksten uit de Nederlandse letterkunde, taalkunde en cultuurgeschiedenis van de vroegste tijd tot heden. De DBNL digitaliseert teksten op een kwalitatief zeer hoog niveau, deze worden volledig opgenomen in een duurzaam XML-TEI-formaat en hebben een foutmarge van minder dan 0,005%. Ook wordt er veel aandacht besteed aan de verrijking van teksten, wat zorgt voor een goede ontsluiting. Er wordt onder andere verwezen naar relevante informatie over auteurs, titels, plaatsnamen en data. Deze handmatig toegevoegde metadata kunnen we gebruiken als trainings- en testmateriaal.

In de volgende stap van onze verkenning bekijken we de mogelijkheid om zowel inhoudelijke als structurele metadata automatisch aan deze teksten toe te voegen. Voor de inhoudelijke metadata willen we gebruik maken van Named Entity Recognition om informatie uit de teksten te halen. Deze informatie moet vervolgens gekoppeld worden aan een thesaurus, waarbij rekening gehouden dient te worden met meervoudige betekenissen (*desambiguatie*) en spellingsvarianten. Voor structurele data gaat het om het aangeven van koppen, paginanummers, gedichten, tabellen en dergelijke. Bij sommige titels (vaak proza) is de opmaak helder, maar voor bijvoorbeeld studieboeken en tijdschriften ligt dit ingewikkelder. De vraag is of we die informatie (gedeeltelijk) automatisch kunnen laten toevoegen.

Met dit onderzoek en met toekomstige verkenningen hopen we een impuls te geven aan de kennis over de mogelijkheden en beperkingen van het automatisch metadateren, nu en in de nabije toekomst. Uitgangspunt daarbij is dat dit het werk van titelbeschrijvers moet vergemakkelijken, maar niet moet overnemen. De menselijke blik, kunde en expertise blijft noodzakelijk om de kwaliteit te waarborgen waar we als KB voor staan.

BRONNEN

1. *Onderzoeksagenda Koninklijke Bibliotheek* 2018-2022. (z.d.). Geraadpleegd van https://www.kb.nl/organisatie/onderzoek-expertise/onderzoeksagenda-2018-2022.

2. Willens, M. (2019, 3 januari). *Forbes is building more AI tools for its reporters.* Geraadpleegd van https://digiday.com/media/forbes-built-a-robot-to-pre-write-articles-for-its-contributors/.

3. Kunova, M. (2018, 6 augustus). Getty Images launches a new AI tool that helps publishers find the right picture for the story. Geraadpleegd van https://www.journalism.co.uk/news/getty-images-launches-a-new-ai-tool-that-helps-publishers-find-the-right-picture-for-the-story/s2/a725797/.

4. He, A. (2018, 9 november). The New York Times Digitizes Millions of Historical Photos Using Google Cloud Technology | The New York Times Company. Geraadpleegd van https://www.nytco.com/press/new-york-times-google-cloud/ & Greenfield, S. (2018, 9 november). *Picture what the cloud can do: How the New York Times is using Google Cloud to find untold stories in millions of archived photos* | Google Cloud Blog. Geraadpleegd van https://cloud.google.com/blog/products/ai-machine-learning/how-the-new-york-times-is-using-google-cloud-to-find-untold-stories-in-millions-of-archived-photos.

5. ICT with Industry 2019 – ICT Research Platform Netherlands. (z.d.). Geraadpleegd van https://ict-research.nl/ict-with-industry/ictwi2019/.

6. ICT with Industry 2019 – ICT Research Platform Netherlands. (z.d.). Geraadpleegd van https://ict-research.nl/ict-with-industry/ictwi2019/ & Sappelli, M., Chu, D., Cambel, B., Graus, D. and Bressers, P. (2018). *Smart journalism: personalizing, summarizing, and recommending financial economic news.* The algorithmic personalization and news (apen18) workshop, The International AAAI Conference on Web and Social Media 2018. Geraadpleegd van https://graus.nu/publications/smart-journalism-position-paper/.

7. Kuiken, J., Schuth, A., Spitters, M. & Marx, M. (2017, 2 februari). *Effective Headlines of Newspaper Articles in a Digital Environment. Digital Journalism*. Geraadpleegd van https://www.tandfonline.com/doi/full/10.1080/21670811.2017.1279978.

8. Kraak, H. (2019, 31 januari). *NPO Start wil je niet meer programma's bieden van jouw smaak, maar je smaak verbreden. Dit is hoe ze dat aanpakken.* De Volkskrant. Geraadpleegd van https://www.volkskrant.nl/cultuur-media/npo-start-wil-je-niet-meer-programma-s-bieden-van-jouw-smaak-maar-je-smaak-verbreden-dit-is-hoe-ze-dat-aanpakken~b59b54c9/?utm_campaign=shared_earned.

9. *KB zet Bookarang in om lezers door te leiden naar volgend boek.* (z.d.). Geraadpleegd van https://www.kb.nl/ob/nieuws/2018/kb-zet-bookarang-in-om-lezers-door-te-leiden-naar-volgend-boek.

10. Thrillseeker. (z.d.). Geraadpleegd van https://thrillseeker.io. & Huijzer, D. (2019, 24 april). *WPG lanceert aanbevelingstool "Thrill Seeker".* Geraadpleegd van https://inct.nl/nieuws/6661/wpg-lanceert-aanbevelingstool-thrill-seeker & Huijzer, D. (2019, 11 juni). *Schwung bij WPG: "Wij leren elke dag nieuwe dingen."* Geraadpleegd van https://inct.nl/news/6749/schwung-bij-wpg-lsquo-wij-leren-elke-dag-nieuwe-dingen-rsquo-.

11. *Talk to Books.* (z.d.). Geraadpleegd van https://books.google.com/talktobooks/ & *Talk to Books by Google AI | Experiments with Google.* (z.d.). Geraadpleegd van https://experiments.withgoogle.com/talk-to-books.

12. van Essen, M. (2019, 27 februari). *Machine Learning en Automatische Classificatie - Evaluatierapport.* Geraadpleegd van https://kia.pleio.nl/groups/view/53406652/kennisplatform-innovatie/blog/view/55809165/machine-learning-en-automatische-classificatie-evaluatierapport.

13. *Tribunaalarchieven als digitale onderzoeksfaciliteit.* (z.d.). Geraadpleegd van https://www.oorlogsbronnen.nl/tribunaalarchieven-als-digitale-onderzoeksfaciliteit & Klijn, E. (2019 april). *Enorme stap voorwaarts om archieven toegankelijk te maken.* IP vakblad voor informatieprofessionals. Geraadpleegd van https://www.oorlogsbronnen.nl/sites/default/files/IP%20mei%202019%20TRIADO%20Edwin%20Klijn.pdf.

14. Hogeweg, L. (2018, 27 maart). *Collaboration started to support biodiversity research through artificial intelligence.* Geraadpleegd van https://science.naturalis.nl/en/about-us/news/onderzoek/collaboration-started-support-biodiversity-research-through-artificial-intelligence/.

15. Speksneijder, C. (2018, 7 september). *Deze slimme camera telt en herkent insecten.* De Volkskrant. Geraadpleegd van https://www.volkskrant.nl/wetenschap/deze-slimme-camera-telt-en-herkent-insecten~bd3b152d/.

16. *Researcher-in-residence.* (z.d.). Geraadpleegd van https://www.kb.nl/organisatie/onderzoek-expertise/researcher-in-residence.

17. Lonij, J., Harbers, F. (2016) *Genre classifier.* (2016). Geraadpleegd van http://lab.kb.nl/tool/genre-classifier. & Broersma, M., Attema, J., Tjong Kim Sang, E. & Klaver, T. (z.d.). *Advancing Media History by Transparent Automatic Genre Classification.* Geraadpleegd van https://www.esciencecenter.nl/project/newsgac.

18. Smits, T., Faber, W.J. (2018) *CHRONReader.* Geraadpleegd van http://lab.kb.nl/tool/chronreader. & Lonij, J., Wevers, M. (2017) *SIAMESE*. Geraadpleegd van http://lab.kb.nl/tool/siamese. & Wevers, M., & Smits, T. (2019). *The visual digital turn: Using neural networks to study historical images.* Digital Scholarship in the Humanities. Geraadpleegd van https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqy085/5296356.

19. Wildschut, P., Faber, W.J. (2017) *Narralyzer*. Geraadpleegd van http://lab.kb.nl/tool/narralyzer.

20. de Jong, A. (2018). *De evolutie van het media asset management bij Beeld en Geluid*. Geraadpleegd van https://publications.beeldengeluid.nl/pub/667.

21. Adrianus J. van Hessen. (z.d.). Geraadpleegd van https://research.utwente.nl/en/persons/adrianus-j-van-hessen & Ordelman, R. J. F., & van Hessen, A. J. (2018). *Speech Recognition and Scholarly Research: Usability and Sustainability*. In I. Skadina, & M. Eskevich (Eds.), CLARIN 2018 Annual Conference (pp. 163-168.) Geraadpleegd van https://research.utwente.nl/en/publications/speech-recognition-and-scholarly-research-usability-and-sustainab. & SpraakLab. (z.d.). Geraadpleegd van https://www.spraaklab.nl/.

22. de Boer, V., Ordelman, R. & Schuurman, (2016) *Evaluating unsupervised thesaurus-based labeling of audiovisual content in an archive production environment.* International Journal on Digital Libraries. Geraadpleegd van https://doi.org/10.1007/s00799-016-0182-6 & de Boer, V., Priem, M., Hildebrand, M., Verplancke, N., de Vries, A., & Oomen, J. (2016). *Exploring Audiovisual Archives Through Aligned Thesauri*. In E Garoufallou, I Subirats Coll, A Stellato, & J Greenberg (Eds.), *Metadata and Semantics Research.* MTSR 2016. Springer, Cham. Geraadpleegd van http://publications.beeldengeluid.nl/pub/632.

23. Junger, U. (2018). *Automation first – the subject cataloguing policy of the Deutsche Nationalbibliothek*. Geraadpleegd van http://library.ifla.org/2213/1/115-junger-en.pdf.

24. *NBD Biblion maakt efficiëntieslag door inzet van kunstmatige intelligentie in samenwerking met Bookarang.* (2018, 19 juni). [Persbericht]. Geraadpleegd van https://www.nbdbiblion.nl/sites/nbdbiblion.nl/files/Persbericht+Samenwerking+Bookarang_def.pdf.

25. Golub, K., Hagelbäck, J., & Ardö, A. (z.d.). *Automatic Classification Using DDC on the Swedish Union Catalogue*. Geraadpleegd van http://ceur-ws.org/Vol-2200/paper1.pdf.

26. Busch, J. (2018). *Automatic Classification Using DDC on the Swedish Union Catalogue*. International Conference on Dublin Core and Metadata Applications. Geraadpleegd van http://dcevents.dublincore.org/IntConf/dc-2018/paper/view/556/669.

27. Brygfjeld, S., Wetjen, F., & Walsøe, A. (2018). *Machine learning for production of Dewey Decimal.* Geraadpleegd van http://library.ifla.org/2216/1/115-brygfjeld-en.pdf.

28. Brygfjeld, S. (2019, 13 juni). *Codename Nancy - AI: lessons from the National Library of Norway* [Slides]. Geraadpleegd van https://www.slideshare.net/sconul/artificial-intelligence-the-national-library-of-norway-svein-arne-brygfjeld-national-library-of-norway.

29. *Annif - tool for automated subject indexing and classification.* (z.d.). Geraadpleegd van http://annif.org & Suominen, O., (2019). *Annif: DIY automated subject indexing using multiple algorithms*. LIBER Quarterly, 29(1), pp.1–25 Geraadpleegd van http://doi.org/10.18352/lq.10285.

30. Google Groepen. (z.d.). [Forum-post]. Geraadpleegd van https://groups.google.com/forum/#!forum/annif-users.

31. Goh, R. (2018). *Using Named Entity Recognition for Automatic Indexing*. Geraadpleegd van http://library.ifla.org/2214/1/115-goh-en.pdf.

32. Hlava, M., Russell, J., & Hansen, D. (2018). *Inverting the Library Cataloguing Process to Streamline Technical Services and Significantly Increase Discoverability and Search for Special Collections*. Geraadpleegd van http://library.ifla.org/2219/1/115-hlava-en.pdf.

33. van Veen, T., Lonij, J., & Faber, W. (2016). *Linking Named Entities in Dutch Historical Newspapers*. Zenodo. Geraadpleegd van http://doi.org/10.5281/zenodo.843504 & van Veen, T. (2019). Wikidata. *Information Technology and Libraries*, 38(2), 72-81. Geraadpleegd van https://doi.org/10.6017/ital.v38i2.10886 & van Veen, T. (2019, 5 april). *Using Wikidata for entity search in historical newspapers* [YouTube]. Geraadpleegd van https://www.youtube.com/watch?v=J5mCem-hEMg.

34. *JSTOR: Text Analyzer.* (z.d.). Geraadpleegd van https://www.jstor.org/analyze/.

35. *"CB distribueert fysieke boeken en e-books, in de winkelstraat en online in Nederland en België. Ook biedt CB logistieke oplossingen voor de Healthcare markt."* (z.d.). Geraadpleegd van https://www.cb.nl/over-ons.

36. *"All ONIX standards are designed to support computer-to-computer communication between parties involved in creating, distributing, licensing or otherwise making available intellectual property in published form, whether physical or digital."* (z.d.). Geraadpleegd van https://www.editeur.org/8/ONIX/.

37. *NBD Biblion levert "een compleet pakket producten en diensten die bijdragen aan het succes van uw bibliotheek of (school)mediatheek"*. NBD Biblion. (z.d.). Geraadpleegd van https://www.nbdbiblion.nl/product/nbd-biblion.

38. Riva, P., Le Bœuf, P., & Žumer, M. (2017). *IFLA Library Reference Model. User Tasks Summary*, p.15. Geraadpleegd van https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf.

39. *"RDA is a package of data elements, guidelines, and instructions for creating library and cultural heritage resource metadata that are well-formed according to international models for user-focussed linked data applications."* About RDA (z.d.). Geraadpleegd van http://www.rda-rsc.org/content/about-rda.

40. Riley, J. (2017). Understanding Metadata. *What is metadata and what is it for?* Geraadpleegd van https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf & Kuipers, A. (2016, 5 juli). *De comeback van de KB-code.* Geraadpleegd van https://www.kb.nl/blogs/over-de-kb/de-comeback-van-de-kb-code & Miller, L. (2010, 9 september). *The trouble with Google Books.* Geraadpleegd van https://www.salon.com/2010/09/09/google_books/.

41. *Geschiedenis van de Nederlandse Bibliografie. (z.d.).* Geraadpleegd van https://www.kb.nl/organisatie/voor-uitgevers/informatie-over-de-nederlandse-bibliografie/geschiedenis-van-de-nederlandse-bibliografie.

42. *Over het Centraal Bestand Kinderboeken.* (z.d.). Geraadpleegd van https://www.kb.nl/bronnen-zoekwijzers/kb-collecties/moderne-gedrukte-werken-vanaf-1801/kinderboeken/over-het-centraal-bestand-kinderboeken.

43. Lorentz Center - *ICT with Industry 2019 from 21 Jan 2019 through 25 Jan 2019.* (z.d.). Geraadpleegd van https://www.lorentzcenter.nl/lc/web/2019/1061/info.php3?wsid=1061.

44. Kleppe, M., Hendrickx, I., Veldhoen, S., Brandsen, A., de Vos, H., Goes, K., ... Zijdeman, R. (z.d.). KB (National Library of the Netherlands): *(Semi-) Automatic Cataloguing of Textual Cultural Heritage Objects.* Geraadpleegd van https://kbresearch.nl/brinkeys/report.pdf.

45. *fastText* (z.d.). Geraadpleegd van https://fasttext.cc.

46. *Annif - tool for automated subject indexing and classification.* (z.d.). Geraadpleegd van http://annif.org.

47. Suominen, O., (2019). *Annif: DIY automated subject indexing using multiple algorithms*. LIBER Quarterly, 29(1), pp.1–25 Geraadpleegd van http://doi.org/10.18352/lq.10285

48.

49. *"TF-IDF is de mate van belang van een specifiek woord in een tekst, dat tot stand komt door het te vergelijken met de frequentie van dat woord in andere teksten."* Groenewoud, R. (2016, 2 november). TF-IDF: *Een nieuwe rage in SEO? - Emerce*. Geraadpleegd van https://www.emerce.nl/best-practice/tfidf-nieuwe-rage-seo.

50. *"Snowball is a small string processing language designed for creating stemming algorithms for use in Information Retrieval."* Snowball. (z.d.). Geraadpleegd van https://snowballstem.org.

51. *Ariadne's Thread: Interactive Context Explorer.* (z.d.). Geraadpleegd van https://www.oclc.org/research/

themes/data-science/ariadne.html & Wang, S. & Koopman, R. (2017), *Clustering articles based on semantic similarity, Scientometrics* (2017) 111: 1017. Geraadpleegd van https://doi.org/10.1007/s11192-017-2298-x.

52. *"In pattern recognition, information retrieval and binary classification, precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Both precision and recall are therefore based on an understanding and measure of relevance in pattern recognition and information retrieval.* (2019). Geraadpleegd van https://en.wikipedia.org/wiki/Precision_and_recall.

53. HathiTrust Digital Library | Millions of books online. (z.d.). Geraadpleegd van https://www.hathitrust.org.

54. *HTRC Analytics.* (z.d.). Geraadpleegd van https://analytics.hathitrust.org.

55. *DBNL · Digitale Bibliotheek voor de Nederlandse Letteren.* (z.d.). Geraadpleegd van https://www.dbnl.org.

KB nationale bibliotheek