Eerste Hulp bij e-Onderzoek voor studenten in de geesteswetenschappen: Slimmer zoeken, slimmer documenteren

Sanders, Ewoud

Citation
Sanders, E. (2011). *Eerste Hulp bij e-Onderzoek voor studenten in de geesteswetenschappen: Slimmer zoeken, slimmer documenteren*. Amsterdam/Den Haag/Leiden: Early Dutch Books Online. Retrieved from https://hdl.handle.net/1887/17774

Version:	Not Applicable (or Unknown)
License:	Leiden University Non-exclusive license
Downloaded from:	https://hdl.handle.net/1887/17774

Note: To cite this publication please use the final published version (if applicable).

Eerste Hulp Bij e-Onderzoek

Ewoud Sanders

Het kolossale aanbod aan digitale bronnen maakt nieuwe onderzoeksvragen mogelijk en vraagt tegelijkertijd om nieuwe manieren van zoeken. Hoe slimmer je zoekt, hoe relevanter de zoekresultaten.
En slimmer documenteren scheelt niet alleen veel tijd, het maakt je onderzoeksresultaten ook herhaalbaar. Slimmer zoeken en documenteren is overigens binnen twee uur te leren: in dit boekje staat hoe.

Uit de tekst:
Opmerkelijk ervaringsfeit: weinig alfa's schrikken van eenvoudige wiskundige formules (op school gehad), maar als ze zoekopdrachten zien als: *komk?????t*d* of *site:dbnl.org "achter de ramen * * tijd"*, dan haken ze al snel af ('dat is mij echt te ingewikkeld'). En dat terwijl je direct, vaak binnen seconden, veel baat kunt hebben bij zo'n geavanceerde zoekopdracht.

Over de auteur: Ewoud Sanders is historicus en journalist. Hij heeft wekelijks een taalrubriek in NRC Handelsblad *en heeft diverse taalboeken op zijn naam staan. Daarvoor maakt hij niet alleen intensief gebruik van internet, maar ook van een grote collectie digitale bronnen in eigen beheer.*

Eerste Hulp Bij e-Onderzoek

voor studenten in de geesteswetenschappen

Slimmer zoeken, slimmer documenteren

Ewoud Sanders

early dutch books ONLINE

Amsterdam/Den Haag/Leiden 2011

Inhoud

1. Inleiding • 3

2. Enkele onmisbare websites • 5

3. Zoeken met operatoren • 13

4. Enkele casestudies • 19

5. The Next Level • 43

Selectie uit de genoemde websites • 48
Belangrijkste zoekoperatoren voor Google en Exalead • 49

1 INLEIDING

Er was een tijd, nog niet eens zo lang geleden, dat het een letterenstudent heel veel tijd kon kosten om zijn of haar onderzoeksgegevens bij elkaar te sprokkelen. Je moest ervoor naar een bibliotheek of archief, soms verspreid over het land, waar je vertrouwd moest raken met kaartenbakken, registers, inventarissen, de microfilmlezer en het kopieerapparaat. Uren, dagen en soms zelfs weken bladeren in oude kranten, tijdschriften of andere bronnen en toch met slechts een klein stapeltje kopieën naar huis gaan – het was geen uitzondering.

Bovendien was het vervolgens een hele klus om de gevonden informatie zo op te bergen dat je er efficiënt mee kon werken. Er zijn talloze scripties, boeken en dissertaties geschreven op basis van kaartenbakken.

Inmiddels is de situatie grondig veranderd. Heel veel instellingen, in binnen- en buitenland, hebben hun collecties via internet toegankelijk gemaakt of zijn daar hard mee bezig. Snel te veel informatie vinden is nu vaker het probleem dan in slakkentempo weinig vinden.

In de meeste bètawetenschappen wordt de ordening van informatie beheerd en geëxploiteerd door commerciële uitgevers. Tegen betaling kun je zoeken in keurig geordende databanken vol artikelen, samenvattingen en onderzoeksresultaten.

In de alfa- of geesteswetenschappen (de wetenschappen die zich bezighouden met de 'geestesproducten' van de mens, waaronder talen, geschiedschrijving, filosofie en kunst), is dat nauwelijks het geval. Dat komt mede doordat het onderzoeksterrein oeverloos is: teksten in alle soorten, maten en smaken, foto's, film, video, muziek, reclame-uitingen – de lijst is lang.

In het recente verleden haalden geesteswetenschappers nog wel eens hun neus op voor digitalisering en voor kwantitatieve methoden, maar de laatste jaren groeit het inzicht dat ook zij geholpen zijn met digitale tools. Daarbij is een centrale rol weggelegd voor *text mining*, kort gezegd: methoden en technieken om automatisch relevante informatie uit min of meer ongestructureerde teksten ('vrije teksten') te halen.

Hoewel er op dit gebied de laatste tien jaar veel vooruitgang is geboekt, valt er nog een hoop te verbeteren. Bovendien zijn de programma's die al wel redelijk goed werken duur en ingewikkeld en dus lang niet voor iedereen toegankelijk. De meeste eerstejaars zien zich vooral geconfronteerd met een kolossale, almaar uitdijende hoeveelheid digitale bronnen. Tientallen miljoenen pagina's aan oude kranten, boeken en tijdschriften bijvoorbeeld – de grootste bibliotheek in de geschiedenis van de mensheid. Die onvoorstelbare hoeveelheid bronnen maakt niet alleen nieuwe onderzoeksvragen mogelijk, maar vereist ook een nieuwe aanpak om die bronnen te bevragen en te doorzoeken, en nieuwe manieren om de gevonden informatie te verwerken.

Dit boekje laat zien hoe je, zonder dure en ingewikkelde programma's, slim kunt zoeken en documenteren. De nadruk ligt op slim zoeken, waarbij we vooral zullen kijken welke mogelijkheden Google biedt. Natuurlijk zijn er ook andere goede zoekmachines, maar Google wordt veruit het meest gebruikt en kan op dit punt het meest, vandaar. Een en ander wordt gepresenteerd aan de hand van acht casussen – onderzoeksvragen zoals studenten die tegenwoordig krijgen voorgelegd.

Aan het eind volgt een korte kennismaking met slim zoeken met behulp van indexeringssoftware in bronnen die je op je eigen pc of laptop hebt gezet. En via zogenoemde *webcrawlers*. Voor sommigen zal dit te ver gaan, voor anderen juist niet. Feit is dat je dan nóg geavanceerder en verfijnder kunt

zoeken dan nu op internet mogelijk is. En heel efficiënt kunt documenteren – met als groot voordeel dat je onderzoek beter verifieerbaar en makkelijker herhaalbaar is.

Tot slot: de afgelopen jaren heb ik geregeld aan (oudere) docenten gevraagd in hoeverre zij vertrouwd waren met slim zoeken op internet. Meestal kreeg ik te horen: 'Ik ben daar zelf niet zo in thuis, maar mijn studenten wel – die kunnen ermee toveren.' Als ik vervolgens aan studenten vroeg in hoeverre zíj vertrouwd waren met slim zoeken op internet ('Ken je bijvoorbeeld deze trucs?'), dan kreeg ik als antwoord: 'Nee, hoe zouden wij die moeten kennen? Die hebben we nooit geleerd. Zelfs niet bij de basisvaardigheden in het eerste jaar.'

Er zullen uitzonderingen zijn, maar ik heb dit antwoord verontrustend vaak gehoord.

Vandaar dit boekje.

Ewoud Sanders

2 ENKELE ONMISBARE WEBSITES

Niemand weet precies hoe groot internet is, maar het is megagroot en je kunt er lelijk verdwalen. Dat kan erg aangenaam zijn – je komt nog eens ergens –, maar ook onpraktisch en tijdrovend. Kennis van goede en betrouwbare sites is onontbeerlijk geworden.

Met betrouwbaar wordt hier bedoeld: de site is gemaakt en wordt onderhouden door een professionele instelling, bijvoorbeeld een universiteit, bibliotheek, museum of archief. Dat wil niet zeggen dat alle informatie die je er vindt altijd voor honderd procent betrouwbaar is, want feiten kunnen achterhaald zijn en bronnen gedateerd. Hier valt veel meer over te zeggen, maar een goede vuistregel is: op websites van professionele instellingen vind je doorgaans de beste informatie om mee van start te gaan. Net als op sites of blogs van experts.

Welke sites dat zijn, hangt af van je studierichting en onderzoek. Bovendien komen er voortdurend sites bij. Het is een kwestie van bijhouden, maar dat hoef je niet alleen te doen. Veel vakgroepen en docenten houden lijstjes bij van betrouwbare websites.

Hieronder een lijstje van websites die elders in dit boekje worden genoemd. Het is een kleine greep – er is zo ontzettend veel –, maar dit zijn enkele gedegen sites waar je heel veel kunt vinden, vooral op het gebied van de Nederlandse en Vlaamse literatuur, taal en (cultuur)geschiedenis. Ook over al deze sites valt meer te vertellen, maar we beperken ons hier tot de hoofdlijnen.

Digitale Bibliotheek voor de Nederlandse Letteren (DBNL)

Webadres: www.dbnl.org

Website over de Nederlandstalige literatuur, taal en cultuurgeschiedenis van de middeleeuwen tot heden. De publicaties worden in twee formaten aangeboden: xml (zeg maar: webpagina's) en pdf. Bij het schrijven van dit boekje telde de site ruim 7500 titels in xml (zo'n 2,3 miljoen webpagina's) en ruim 17.500 pdf's. Maandelijks komen er zo'n 50.000 webpagina's bij. De site bevat literaire teksten, secundaire literatuur en aanvullende informatie zoals biografieën en portretten. Let op: de zoekmogelijkheden bij de DBNL lijken erg geavanceerd, maar laten te wensen over (zie als voorbeeld de vierde case-study).

Woordenboek der Nederlandsche Taal

Webadres: http://gtb.inl.nl/?owner=WNT

Het *Woordenboek der Nederlandsche Taal* is het belangrijkste historische woordenboek voor het Nederlands. De eerste aflevering verscheen in 1864, de laatste aanvullingen in 2001 (productietijd: 137 jaar!). Het telt zo'n 400.000 trefwoorden en beschrijft, mede op basis van ruim 1,7 miljoen citaten, het Nederlands van grofweg 1500 tot 1976. Het WNT maakt deel uit van de Geïntegreerde Taalbank (GTB) van het Instituut voor Nederlandse Lexicologie. Tegelijk met het WNT kun je enkele andere grote wetenschappelijke woordenboeken doorzoeken: het Oudnederlands Woordenboek, het Vroegmiddelnederlands Woordenboek, het Middelnederlandsch Woordenboek en het Woordenboek der Friese taal. Het kost tijd om goed met deze bronnen te werken, maar ze bevatten een schat aan informatie.

Etymologiebank

Webadres: www.etymologiebank.nl

Of je nu onderzoek doet naar de komma (bestaat al een monografie over) of naar de invloed van het cartesiaans dualisme in de hedendaagse Westerse filosofie, het is altijd nuttig om de geschiedenis van een begrip te bekijken. Daarvoor is de Etymologiebank een onmisbare bron (etymologen onderzoeken de oorsprong en geschiedenis van woorden en uitdrukkingen). Op deze site zijn ruim twintig etymologische woordenboeken op woordniveau ontsloten, inclusief het vierdelige *Etymologisch woordenboek van het Nederlands* (verschenen tussen 2003 en 2009). De bronnen worden in chronologische volgorde gepresenteerd – van jong naar oud – en er komen regelmatig nieuwe bij. Gekoppeld aan de Etymologiebank is de Etymologiewiki: daarin kan iedereen correcties en aanvullingen op de Etymologiebank publiceren.

Historische kranten

Webadres: http://kranten.kb.nl

Kranten geven de eerste ruwe schets van de geschiedenis en zijn daarom belangrijke bronnen. Om die reden worden de laatste jaren heel veel oude kranten gedigitaliseerd: alleen al in Nederland zijn hier ruim honderd instellingen en archieven mee bezig. De grootste, helderste en best doorzoekbare krantensite is die van de Koninklijke Bibliotheek, Historische Kranten genaamd. Er zijn nu 1,5 miljoen pagina's doorzoekbaar van kranten uit de jaren 1618-1945 (waaronder duizenden kranten en krantjes uit de Tweede Wereldoorlog). Bij de voltooiing van de eerste fase van dit project, naar verwachting eind 2012, zullen er ruim 9 miljoen krantenpagina's doorzoekbaar zijn, tot 1995. Het gaat niet om *alle* kranten die ooit in Nederland zijn uitgegeven (misverstand), maar om een *selectie* van landelijke, regionale, lokale, Nederlands-Indische, Surinaamse en Antilliaanse kranten.

Voor een nuttig overzicht van andere gedigitaliseerde kranten uit Nederland, Vlaanderen (en diverse andere landen), zie: www.kranten-historisch.startpagina.nl.

De Handelingen

Webadres: www.statengeneraaldigitaal.nl

Aanvankelijk werden de beraadslagingen van de leden van de Eerste en de Tweede Kamer samengevat in notulen, maar vanaf 1849 werden ze bijna woordelijk opgeschreven door stenografen (pas sinds juni 2001 wordt 'onparlementair taalgebruik' ook opgenomen, daarvoor werden beledigingen en scheldpartijen gecensureerd). Alle Kamerverslagen, Kamerstukken en Kamervragen kwamen terecht in de zogenoemde *Handelingen*, kolossale boeken die in bibliotheken tientallen meters boekenplank in beslag namen. De complete Handelingen van 1814 tot 1 januari 1995 zijn inmiddels digitaal doorzoekbaar bij

www.statengeneraaldigitaal.nl. Vanaf 1995 zijn ze beschikbaar via de website Overheid.nl, bij https://zoek.officielebekendmakingen.nl/zoeken. De Handelingen zijn niet alleen een onmisbare bron voor de parlementaire geschiedenis, maar ook zeer interessant voor onderzoek naar normen en waarden, taalontwikkelingen, enzovoorts.

Google Books

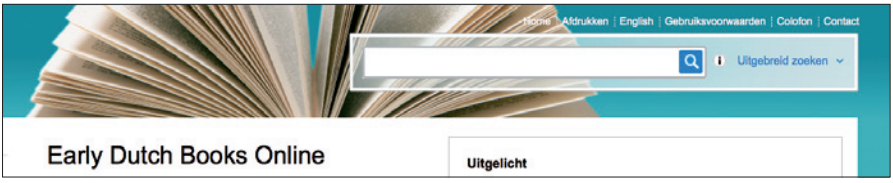
Webadres: www.books.google.nl of www.books.google.com

In 2004 begon Google met een project dat nu bekendstaat als Google Books of Google Book Search. Ambitie: zo'n beetje alle boeken ter wereld digitaal ontsluiten, te beginnen met de boeken die rechtenvrij zijn (om nog meer rechtszaken over het auteursrecht te voorkomen beperkt Google zich nu voornamelijk tot boeken van vóór 1871). In 2007 ging Google een samenwerking aan met de Universiteitsbibliotheek Gent (300.000 boeken tot 1870), in 2010 met de Koninklijke Bibliotheek in Den Haag (160.000 boeken van 1700 tot 1870). Er is veel kritiek mogelijk op Google Books, maar het is massa, bulk, een kolossale boekenberg die je niet over het hoofd kan of mag zien. Er zijn nu zo'n 15 miljoen boeken gedigitaliseerd (de meeste Engelstalig); het worden er zo'n 130 miljoen.

Zeker ook de moeite waard is Google Scholar (http://scholar.google.nl), een enorme verzameling wetenschappelijke artikelen, samenvattingen, proefschriften en boeken van academische uitgeverijen, universiteiten en andere wetenschappelijke organisaties.

Early Dutch Books Online

Webadres: www.earlydutchbooksonline.nl

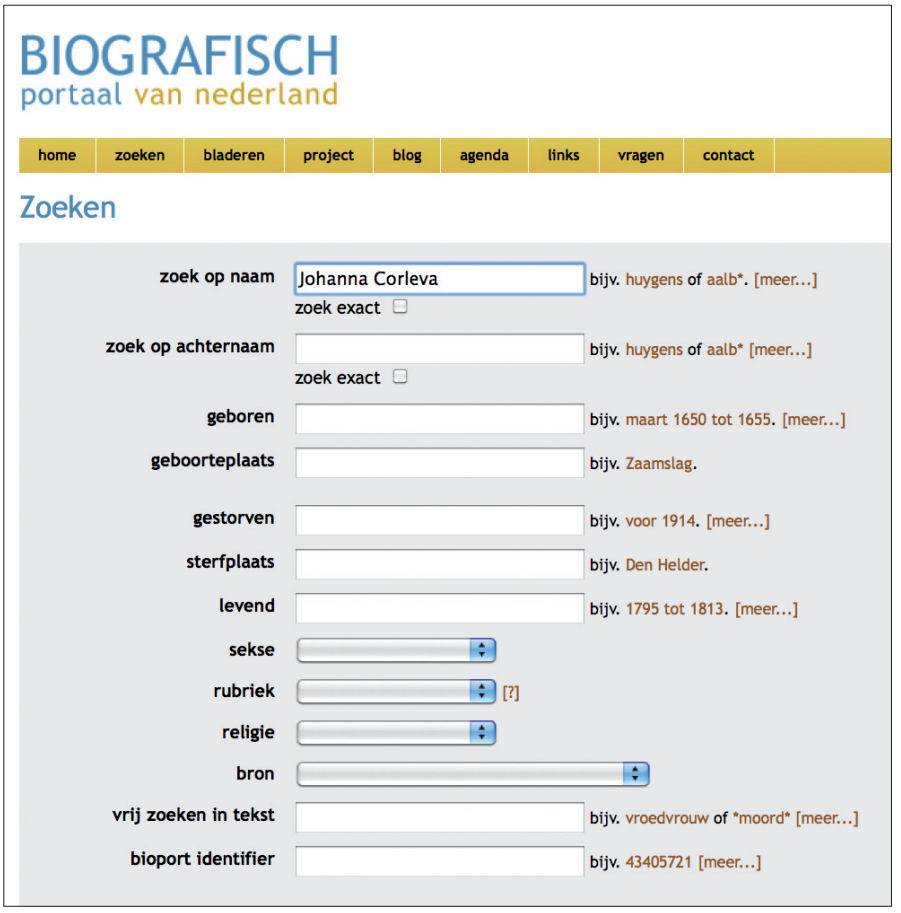


Het is de bedoeling dat over enkele decennia alle boeken uit en over Nederland digitaal beschikbaar zijn, maar zoiets gaat stapsgewijs. In 2011 werd een stap gezet met de lancering van Early Dutch Books Online (EDBO) – een collectie van 11.240 boeken uit de periode 1781-1800. De boeken zijn geselecteerd

uit de STCN (Short Title Catalogue Netherlands) en beslaan diverse thema's: geschiedenis en politiek, economie en handel, wetenschap en ideeëngeschiedenis, godsdienst en samenleving, dagelijks leven en onderwijs, beeldende kunst, literatuur en toneel. In totaal gaat het om ruim 2,1 miljoen pagina's (waarvan acht procent in het Frans), die op woordniveau doorzoekbaar zijn. De boeken, van pamfletjes van vier pagina's tot dikke pillen, zijn pagina voor pagina te bekijken (in kleur), maar ook te downloaden als pdf (in z/w).

Biografisch Portaal van Nederland

Webadres: www.biografischportaal.nl



Op het Biografisch Portaal van Nederland zijn nu 20 biografische bronnen ontsloten, met daarin korte levensschetsen van ruim 70.000 mensen. De website is continu in ontwikkeling en groeit voortdurend. Het is de bedoeling dat op het Biografisch Portaal uiteindelijk over iedereen die in het Nederlandse taalgebied ooit van enige betekenis is geweest, informatie te vinden is. Er zijn links naar archivalia van het Nationaal Archief, en uitgebreide zoekmogelijkheden. De site wordt onderhouden door het Huygens ING, het grootste Nederlandse instituut voor de geesteswetenschappen, waar nog veel meer interessante bronnen te vinden zijn (zie www.huygens.knaw.nl).

Meertens Instituut

Webadres: www.meertens.knaw.nl

Het Meertens Instituut is een onderzoeksinstituut dat zich bezighoudt met de bestudering en documentatie van de Nederlandse taal en cultuur. De website bevat allerlei interessante databanken, over familienamen en voornamen bijvoorbeeld, over oude Nederlandse maten en gewichten, over volkskundige boeken en artikelen, en over ruim 125.000 Nederlandse liederen, van de middeleeuwen tot de moderne tijd (inclusief liefdesliederen, spotliederen, religieuze liederen, sinterklaas- en kerstliedjes).

Het Geheugen van Nederland

Webadres: www.geheugenvannederland.nl

Tot slot Het Geheugen van Nederland, een website die – bij het samenstellen van dit boekje – 553.135 items uit 102 collecties van 90 instellingen bevatte. Je kunt er op foto's, video's, geluidsopnames en teksten zoeken.

Gebruikersvriendelijk?

Doen we met deze selectie allerlei bibliotheken, archieven en andere professionele instellingen tekort? Zeker, want er zijn natuurlijk nog veel meer nuttige en interessante websites. Sommige stadsarchieven hebben prachtige sites met rijke collecties, bijvoorbeeld:

- **www.stadsarchief.amsterdam.nl**
- **www.hetutrechtsarchief.nl**

Op de website van de Universiteit van Leiden is een geweldige collectie van ruim 11.000 toneelteksten tot 1803 te vinden.

- **www.let.leidenuniv.nl/Dutch/Ceneton**

Er zijn belangrijke catalogi; de Bibliografie van de Nederlandse Taal- en Literatuurwetenschap en de Digitale Bibliografie Nederlandse Geschiedenis bijvoorbeeld:

- **www.bntl.nl**
- **www.dbng.nl**

Tijdschriften als *De Groene Amsterdammer* en *Elsevier's Geïllustreerd Maandschrift* hebben zelf hun interessante historische archieven online gezet:

- **www.groene.nl**
- **www.elseviermaandschrift.nl**

En zo gaat het maar door – het is soms om moedeloos van te worden.

Belangrijker is echter de vraag: zijn de sites zo ontsloten dat je ze makkelijk kunt doorzoeken? Zijn ze gebruikersvriendelijk?

Dat verschilt sterk. Bij sommige sites wordt helder uitgelegd wat de eenvoudige en geavanceerde zoekmogelijkheden zijn en bij de Historische Kranten is zelfs een instructiefilmpje te bekijken. Andere sites beloven allerlei zoekmogelijkheden maar maken die (nog) niet waar. En soms kun je veel slimmer zoeken dan op de website wordt verteld.

Kortom, het wordt de student en onderzoeker niet makkelijk gemaakt, maar dat hoort bij nieuwe technologie. Zoals je vroeger in bibliotheken en archieven vertrouwd moest raken met inventarissen, de microfilmlezer en het kopieerapparaat, zo moet je nu handig zien te worden met de zoekmogelijkheden op websites.

Bij websites die bekend zijn bij grote zoekmachines zoals Google, kun je je die moeite soms beter besparen. Via zoekmachines kun je websites namelijk benaderen met zogenoemde *zoekcommando's* of *operatoren*. Dat kan niet alleen veel tijd schelen, maar ook betere informatie opleveren.

Kennis van goede en betrouwbare sites is onontbeerlijk geworden. Sites doorzoeken kan echter een hele klus zijn. Ze werken allemaal een beetje anders, soms kunnen ze minder dan ze beloven en soms juist meer. Websites die worden geïndexeerd door de grote zoekmachines kun je soms beter op een andere manier benaderen: met slimme zoekcommando's.

Ocr en wildcards

Je kunt boeken, kranten en tijdschriften digitaliseren door ze over te (laten) tikken, maar omdat dit veel tijd en geld kost worden ze meestal pagina voor pagina gescand of gefotografeerd. Om de teksten op de afbeeldingen vervolgens toch doorzoekbaar te maken, worden ze door de computer gelezen met behulp van zogeheten *optical character recognition* (optische tekenherkenning, afgekort ocr).

Is het belangrijk om dat te weten? Ja, want bij het ocr'en kan er van alles misgaan. In het algemeen kun je zeggen: hoe ouder de tekst, hoe slechter de kwaliteit van de ocr. Bij teksten van grofweg voor 1860 stijgt het foutenpercentage snel. Je kunt vaak nog veel in dergelijke teksten vinden, maar dan moet je bij het zoeken rekening houden met veelvoorkomende ocr-fouten.

Zo wordt *sch* in oude teksten door ocr-software vaak herkend als *fch*, omdat de s in oud drukwerk sterk leek op de f. Wie zoekt naar *wetenschap*, doet er in dergelijke oude bronnen dus goed aan om ook op *wetenfchap* te zoeken.

Gelukkig is het in sommige databanken – zoals bij de Historische Kranten – mogelijk om te zoeken met behulp van zogenoemde wildcards of jokertekens. De bekendste zijn: ? en *.

Met het vraagteken vervang je één willekeurig teken. Met *weten**?**chap* vind je dus zowel *weten**s**chap* als *weten**f**chap*. Met een sterretje of asterisk vervang je een willekeurig aantal tekens. Met *wetenschap** vind je dus ook woorden als *wetenschapper(s)* en *wetenschappelijk(e)*.

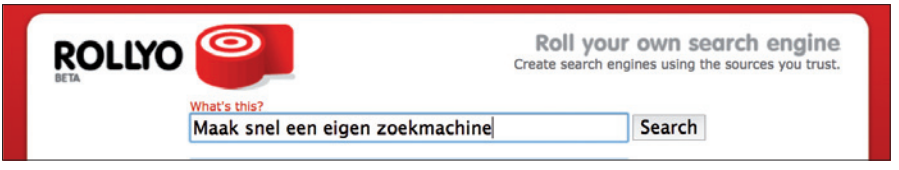
Je kunt deze jokertekens ook combineren: *weten?chap** bijvoorbeeld.

Helaas hanteren niet alle databanken dezelfde jokertekens. Wat dat betreft is het handig om altijd even de toelichting bij 'geavanceerd zoeken' te bekijken. En het loont de moeite om gewoon dingen te proberen: soms kan er meer dan er wordt verteld.

Veel wetenschappers zien ocr-fouten als de achilleshiel van de massadigitalisering ('door al die fouten weet je nooit wat je allemaal mist!'). Ze hebben daar gelijk in, maar door geduldig en creatief met wildcards te zoeken, kun je wel het een en ander compenseren. Meer hierover in casestudy 7.

Maak je eigen zoekmachine

Werk je, op zoek naar informatie, vaak een vast lijstje met websites af? Met telkens dezelfde zoekopdracht? Dat kan handiger. Er zijn allerlei websites waar je heel snel een eigen zoekmachine kunt maken. Je kunt er je vaste lijst met websites opgeven en bij de volgende zoekopdracht doorzoek je in één keer die tien, twintig of meer geselecteerde websites. Kijk bijvoorbeeld eens bij www.rollyo.com of zoek naar 'Google aangepast zoeken'. Het heeft overigens geen zin om in zo'n lijst websites op te nemen met databanken die je alleen rechtstreeks mag doorzoeken.



3 ZOEKEN MET OPERATOREN

We kennen het allemaal: je zoekt iets op internet en je krijgt honderden, soms zelfs duizenden resultaten. Ga je die allemaal bekijken? Nee, doorgaans beperken we ons tot de eerste tien of twintig hits.

Maar waren dat wel de meest relevante? Zoals bekend kom je bij Google hoger op de resultatenlijst als je hiervoor betaalt – daar verdient de zoekmachine haar geld mee. Wie bijvoorbeeld intikt: **boeken Willem Bilderdijk** (aantal resultaten: 129.000), krijgt bovenaan twee verwijzingen naar bol.com (en dit zijn géén advertenties). Pas later volgt een verwijzing naar de Digitale Bibliotheek voor de Nederlandse Letteren (DBNL), hoewel daar bijna honderd boeken van veelschrijver Bilderdijk in digitale vorm worden aangeboden, plus een uitvoerige bibliografie, afbeeldingen, tientallen publicaties over Bilderdijk en verwijzingen naar relevante sites, waaronder het Bilderdijkmuseum.

Nou stond in dit geval de verwijzing naar de DBNL al op de veertiende plaats, dus de meeste mensen zouden 'm wel gevonden hebben, maar je kunt wel degelijk nuttige zoekresultaten over het hoofd zien doordat ze te laag op de lijst staan. In het algemeen kun je zeggen: als je lang moet doorklikken voordat je hebt gevonden wat je zocht, dan was je zoekopdracht niet exact genoeg.

Om preciezer te zoeken kun je gebruikmaken van zogenoemde zoekoperatoren of zoekcommando's. Dit zijn woorden of tekens die je aan een zoekopdracht toevoegt – gewoon in de zoekregel van een zoekmachine (vanaf nu: Google). Er bestaan tientallen van dergelijke zoekoperatoren. Een compleet en helder overzicht ontbreekt helaas, maar voor een aardig overzicht zie:

- **http://tinyurl.com/lijst1-operators**
- **http://tinyurl.com/lijst2-operators**

Ook in de helpfunctie van Google is het nodige te vinden, zij het minder dan je zou verwachten.

Hieronder volgt een lijstje van zoekoperatoren waar je bij letterkundig, taalkundig en historisch onderzoek veel aan kunt hebben, met een korte uitleg. Sommige operatoren zien er bedrieglijk simpel uit – je zou er bijna je schouders over ophalen. In de praktijk kunnen ze echter, net als wildcards, indrukwekkend krachtig zijn, trefzekere stuurlui in een oceaan van informatie.

Als je een beetje handig wordt met dit soort zoekoperatoren, zul je al snel zien dat je resultatenlijst krimpt, terwijl de relevantie van de resultaten met grote sprongen stijgt. Dat lukt overigens zeker niet altijd na één zoekopdracht: slim zoeken is creatief werk. Je probeert nu eens deze truc, dan die. Wees gewaarschuwd: voor je het weet krijg je er zelfs lol in. (En voor studenten die een baan bij een uitgeverij of in de media ambiëren: slim kunnen zoeken heeft marktwaarde, bij redacties en documentatieafdelingen is er veel vraag naar.)

In het volgende hoofdstuk gaan we uitvoeriger met zoekoperatoren aan de slag aan de hand van enkele casussen. Hier vast een beknopte uitleg. Plus, voor de duidelijkheid, een of twee korte voorbeelden. Om misverstanden te voorkomen, je vult die zoekoperatoren in bij de gewone/standaardversie van Google, dus niet bij Google Geavanceerd Zoeken.

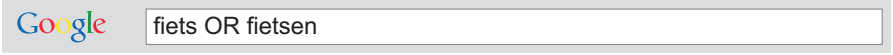
De booleaanse zoekoperatoren

AND, OR, NOT worden booleaanse zoekoperatoren genoemd. Ze danken hun naam aan de Britse wiskundige George Boole (1815-1864), achteraf beschouwd als een van de grondleggers van de computerwetenschap. AND (te vervangen door het plusteken) is een standaardinstelling bij Google. De zoekopdracht **George +Boole** levert dus hetzelfde op als **George Boole**. Toch kan de zoekoperator + wel degelijk van pas komen.

AND [of +] = precies zo moet het er staan; dit moet voorkomen
Voorbeeld:

Hiermee zeg je: er moet echt *tikkfout* staan. Google wil dit corrigeren ('bedoelde u: +tikfout?'), maar nee, dat bedoelde je dus niet. Google negeert veelvoorkomende woorden als *de*, *het*, *waar* en *hoe* (zogenoemde *stopwords*), maar als zo'n woord essentieel is voor je zoekopdracht, dwing je Google om ernaar te zoeken door + voor het woord te zetten. Let op: voor + moet een spatie staan, erna niet. Als je AND gebruikt in plaats van + dan moet AND met hoofdletters.

OR = geef me het ene OF het andere woord
Voorbeeld:



Met AND zeg je: precies zo moet het er staan. Met OR zeg je: laat me websites zien waar *of* dit *of* dat staat: relatie *of* verkering, studiefinanciering *of* stufi, enzovoorts. Vaak is het essentieel om zowel het enkelvoud als het meervoud van een woord in te tikken: *depressie* OR *depressies*. Je kunt met OR ook goed zoeken naar spellingvarianten van woorden of namen (*huygens* OR *huijgens*), naar varianten in andere talen (*sjostakovitsj* OR *shostakovich* OR *chostakovitch*), enzovoorts. Ervaring leert dat als je maar op één woordvorm zoekt (alleen enkelvoud bijvoorbeeld), je waardevolle bronnen kunt missen.
Let op: OR moet met hoofdletters.

NOT [of -] = sluit dit woord uit
Voorbeeld:

Hiermee vind je van alles, maar niet *New York*.
Let op: voor - komt wel een spatie, erna niet. Als je in plaats van - NOT gebruikt, dan moet dit met hoofdletters.

Puntjes, sterretjes en meer

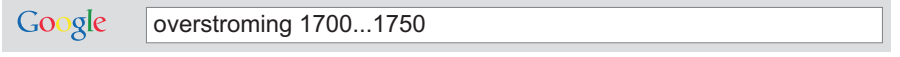
" " [dubbele aanhalingstekens] = geef me exact deze woordgroep, in deze volgorde
Voorbeeld:

Let op: dit werkt niet met *enkele* aanhalingstekens.

***** [asterisk of sterretje] = vervangt ieder willekeurig woord
Voorbeeld:

Resultaat: niet alleen *zo sterk als een beer* en *zo sterk als een paard*, maar ook allerlei varianten hierop. Over het * zéker niet de schouders ophalen, want bij slim zoeken is dit een machtig wapen.
Let op: in het vorige hoofdstuk kwam ter sprake dat je in *databanken*, bijvoorbeeld in de Historische Kranten, binnen een woord een willekeurig aantal tekens kunt vervangen door een asterisk. Bij de grote zoekmachines op internet is dat niet het geval. Daar vervang je met een * een *heel* woord, niet een woord*deel*. Voor en na het sterretje moet dan ook een spatie staan, anders werkt het niet. Voor de (beperkte) uitzondering op deze regel, zie het kader 'Exalead' in het volgende hoofdstuk.

... [drie puntjes] = geef me een waarde tussen x…y
Voorbeeld:



Resultaat: websites waarop sprake is van een overstroming tussen 1700 en 1750. Werkt (natuurlijk) ook met bedragen, bijvoorbeeld: laptop 100...150 euro.
Let op: geen spaties voor of na de puntjes plaatsen. Werkte voorheen alleen met drie puntjes, maar tegenwoordig ook (vaak) met twee.

~ [tilde] = zoek ook naar synoniemen
Voorbeeld:

Resultaat: naast *violence* ook *abuse*, *crime stories*, enzovoorts.
Werkt beter voor het Engels dan voor het Nederlands.

Enkele handige zoekoperatoren

allintitle: = alle gezochte woorden moeten in de titel van de webpagina voorkomen
Voorbeeld:

'In de titel van de webpagina' betekent: prominent op de pagina, meestal ergens bovenaan en in grotere letters. Mocht je slechts één 'titelwoord' zoeken, dan kun je volstaan met de zoekoperator **intitle:**
Let op: geen spatie na de dubbele punt.

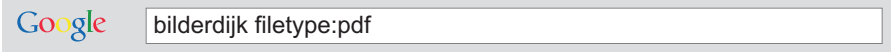
allinurl: = alle gezochte woorden moeten in het adres (de url) van de webpagina voorkomen
Voorbeeld:

Mocht je slechts één woord in een url zoeken, dan kun je volstaan met de zoekoperator **inurl:**
Let op: geen spatie na de dubbele punt.

define: = geef de definitie van het gezochte woord
Voorbeeld:

Werkt beter voor het Engels dan voor het Nederlands.
Let op: geen spatie na de dubbele punt.

filetype: = zoek dit bestandsformaat
Voorbeeld:



Als je eenmaal weet dat de meeste rapporten, boeken, scripties en proefschriften in pdf-formaat op internet worden gezet (reden: je kunt er niet makkelijk iets in wijzigen), dan is het handig om dat bestandsformaat aan je zoekopdracht toe te voegen. Je kunt op duizenden bestandsformaten zoeken (zie voor een overzicht www.file-extensions.org). Meestal kom je al een heel eind met deze:

doc of **docx** = MS Word (veel gebruikt formaat voor notulen, vergaderstukken e.d.)
pdf = Portable Document Format
ppt of **pptx** = PowerPoint (veel gebruikt voor presentaties)
kmz = Google Earth
xls of **xlsx** = Excel (veel gebruikt voor berekeningen, begrotingen e.d.)

Let op: geen spatie na de dubbele punt en haal de punt voor de extensie weg (dus niet **filetype:.pdf**, maar **filetype:pdf**). Of je eerst de operator **filetype:** intikt en daarna de zoekterm(en), of andersom, maakt meestal niet uit.

site: = beperk de zoekopdracht tot deze website of dit domein

Voorbeeld:

Bovenste resultaat: het colofon van de redactie van *NRC Handelsblad*.

Je kunt ook aangeven dat je alleen in websites van een bepaald land wilt zoeken: bijvoorbeeld alleen uit Nederland (.nl), België (.be), Duitsland (.de), Frankrijk (.fr) of het Verenigd Koninkrijk (.uk). Er bestaan honderden van dergelijke topleveldomeinen, zoals ze officieel heten. Voor een overzicht, zie het artikel 'Lijst van topleveldomeinen op het internet' in Wikipedia.

Let op: geen spatie na de dubbele punt. En anders dan de helpfunctie van Google aangeeft, hoef je de topleveldomeinen niet vooraf te laten gaan door een punt. Dus niet **site:.nl** of **site:.be**, maar simpelweg **site:nl** of **site:be**

Enkele andere nuttige topleveldomeinen:

edu = instellingen voor hoger onderwijs, o.a. Amerikaanse universiteiten
eu = Europese Unie
gov = Amerikaanse instellingen van de federale overheid
int = internationale organisaties (handig voor internationale verdragen)
museum = inderdaad, musea
org = non-profit organisaties

Belangrijk om te weten: *je vindt vaak meer op een site als je die via Google benadert, dan via het zoekvenster op zo'n site*. Dat lijkt vreemd – bizar zelfs –, maar is wel te verklaren: Google kan veel meer geld en moeite stoppen in optimale indexering (hun *core business*) dan de meeste websites, zelfs als die groot zijn. Zo kun je bij de Digitale Bibliotheek voor de Nederlandse Letteren (DBNL) allerlei gegevens wél via Google vinden, maar niet rechtstreeks. Websites benaderen met de site-operator vanuit Google heeft als bijkomend voordeel dat je je niet telkens in wéér een andere gebruikersinterface hoeft te verdiepen. Zoals al eerder gezegd: lang niet alle databanken op internet geven Google (en andere zoekmachines) toestemming om hun content te indexeren. Het is kwestie van uitproberen.

Er zijn nog veel meer zoekoperatoren. Hier nog een paar die nuttig kunnen zijn (afhankelijk van wat je zoekt):

discussions: = zoek in discussiegroepen, opiniepagina's en vraag-en-antwoordpagina's
inblogtitle: = zoek in de titels van blogs
loc: = zoek alles over deze locatie
movie: = zoek informatie over films
related: = zoek verwante, vaak buitenlandse sites

Zoekoperatoren combineren

Soms is het voldoende om gebruik te maken van slechts één zoekoperator, maar vaker zul je zoekoperatoren combineren en/of zoekoperatoren herhalen met een andere waarde. Ik zoek dit, maar het kan ook zo gespeld worden (OR), of zo (OR), dit moet niet in de resultaten staan (-), dit trouwens ook niet (-), en voor de zekerheid laat ik binnen een formulering een paar dingen open (**).

Niet gevonden wat je zocht? Nog steeds te veel ruis of vindt Google juist helemaal niks? Dan probeer je het gewoon opnieuw, op een iets andere manier. Bij iedere poging houd je de zoekvraag overigens rechttoe rechtaan, want je praat met een computer, niet met een mens. Hoofdletters doen er niet toe (behalve dan bij de booleaanse operatoren) en het is handig om van tevoren te bedenken hoe iets in een document zou kunnen staan ('Denk als een document', is het advies van Henk van Ess, een van de beste internetzoekers van Nederland).

Zou jij zelf boven een lijst met boeken van Willem Bilderdijk **boeken van Willem Bilderdijk** zetten? Dat zou kunnen, maar mij lijkt het veel waarschijnlijker dat je het woord *bibliografie* gebruikt – want dat is in de wetenschap veel gangbaarder. Dat maakt de zoekopdracht **bibliografie Willem Bilderdijk** of **bibliografie Bilderdijk** slimmer. Zelfs zonder zoekoperatoren maakt dit de eerste zeven resultaten echt relevant, en bol.com verhuist naar de achtste plaats.

Welkom in de wereld van het slimme zoeken. Waar minder meer is.

Voor een samenvatting van de belangrijkste zoekoperatoren, zie de achterflap van dit gidsje.

Slim zoeken is een sport. Voornaamste hulpmiddelen: zoekoperatoren. Dit zijn woorden of tekens die je aan een zoekopdracht toevoegt – gewoon in de zoekregel van een zoekmachine. Gevolg: je resultatenlijst krimpt, terwijl de relevantie van de resultaten met grote sprongen stijgt.

4 ENKELE CASESTUDIES

Hierna volgen enkele casestudies. De opdrachten zijn opzettelijk beknopt gehouden. In het echte leven zou zo'n opdracht een stuk gedetailleerder zijn. Hij zou dan bijvoorbeeld kunnen luiden (deze tekst komt van een universitair docent):

> In deze opdracht ga je na hoe je informatie kunt vinden over dit onderwerp en op welke manier de opgegeven sites geschikt zijn als instrumenten voor het opsporen van wetenschappelijke informatie. Het gaat in deze opdracht vooral om de evaluatie van het zoeken en van het zoekresultaat.
>
> Selecteer een onderwerp. Verzamel over dit onderwerp informatie via de opgegeven sites. Noteer je bevindingen. Stel daarbij in ieder geval de volgende zaken aan de orde:
> - Hoe heb je gezocht en hoeveel informatie leverde dit op?
> - Hoe kun je de wetenschappelijke kwaliteit van de gevonden informatie beoordelen? Hoe *controleerbaar* en *actueel* is de informatie?
> - Hoe kun je de relevante informatie die je zoekactie heeft opgeleverd scheiden van de niet-relevante?

Heel nuttig allemaal, maar de onderstaande opdrachten hebben een ander doel: laten zien welke trucs je kunt toepassen om de beste informatie te vinden, liefst snel. De meeste casestudies hebben betrekking op Nederlandse taal- en letterkunde, maar de beschreven zoektechnieken kun je natuurlijk veel breder inzetten. Het spreekt voor zich dat je je *zoektaal* moet aanpassen als je bijvoorbeeld naar informatie uit Latijns-Amerika of China zoekt.

Koninklijke Bibliotheek

Een gouden tip voor wie niet is verbonden aan een universiteit: word lid van de Koninklijke Bibliotheek (huidige tarief: € 15 per jaar). Dit geeft je niet alleen toegang tot PiCarta en WorldCat, maar tot een schat aan nationale en internationale digitale bestanden (catalogi en fulltext-bestanden).

Enkele verborgen schatten in de digitale collecties van de KB.

Casestudy 1

Zoek publicaties over een bepaald onderwerp. Beargumenteer waarom jij denkt dat deze publicaties betrouwbaar zijn. In dit geval: zoek een recente Nederlandse publicatie over racisme en integratie.

De belangrijkste Nederlandse catalogus is PiCarta. PiCarta is een zogenoemde meta-catalogus, die toegang biedt tot de Nederlandse Centrale Catalogus (NCC) en Online Contents. De NCC bevat titelbeschrijvingen van alle boeken en tijdschriften die ruim 400 Nederlandse bibliotheken in hun bezit hebben. Online Contents bevat gegevens over tijdschriftartikelen (vanaf 1992) en wordt samengesteld op basis van de inhoudsopgaven van bijna 15.000 tijdschriften op alle wetenschapsgebieden.

PiCarta is een kolossale catalogus. In 2007 zaten er 64 miljoen titelbeschrijvingen in, momenteel ruim 70 miljoen. Het loont de moeite, zeker aan het begin van je studie, om met deze catalogus vertrouwd te raken. Dat kun je doen door dingen uit te proberen of door een keer de handleiding te lezen.

Hoe vind je een recente handleiding? Probeer eens:

Tussen de eerste tien resultaten vind je:

Wie opziet tegen het lezen van zo'n lange handleiding, vindt met de zoekopdracht:

een nuttige interactieve handleiding.

Goed, met PiCarta zul je nog vaak genoeg werken, het is een onmisbare catalogus. Nadeel is dat je er alleen titelbeschrijvingen vindt, geen digitale boeken of artikelen (wel vind je er steeds vaker een link naar het volledige boek of artikel, elders op internet). Hier willen we verder op zoek naar literatuur waar je meteen mee aan de slag kunt, naar *content* dus.

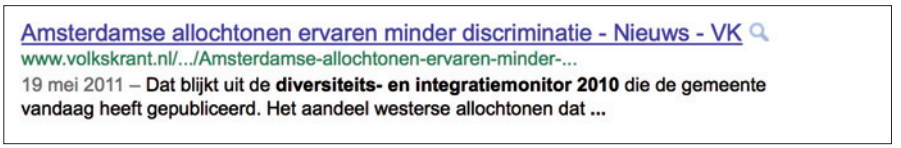
Zoals eerder gezegd kan de zoekoperator **filetype:** wat dit betreft goede diensten bewijzen. Aangezien veel artikelen, boeken, scripties en dissertaties in pdf-formaat op internet worden gezet, beginnen we met de zoekopdracht:

Dit geeft echter zoveel hits dat we er twee dingen aan toevoegen: we willen alleen zoeken op Nederlandse websites en alleen publicaties van 2010 of later. De zoekopdracht wordt nu dus:

Dat geeft als eerste resultaat:

Het blijkt een rapport van 145 pagina's te zijn, getiteld *Diversiteits- en integratiemonitor 2010* van de Dienst Onderzoek en Statistiek van de Gemeente Amsterdam. In de voetnoten staan tientallen verwijzingen naar andere recente literatuur. Een krachtig argument dat het hier om een betrouwbare publicatie gaat, lijkt mij dat dit rapport afkomstig is van een serieuze dienst bij een grote gemeente.

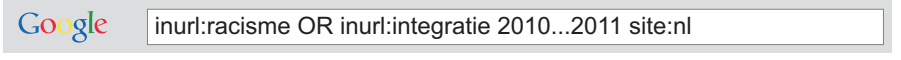
Snel nog even kijken hoe enkele kranten dit rapport bij verschijning hebben samengevat of ontvangen:



En:

Voor de goede orde: niet alle archieven van kranten zijn openbaar of gratis toegankelijk.

Natuurlijk moet je ook altijd kijken of de woorden die je zoekt voorkomen in url's. In dit geval zou de zoekopdracht er zo uit kunnen zien:



Ook dat levert meteen nuttige resultaten op, op betrouwbare sites, bijvoorbeeld:

Casestudy 2

Een bron die je gebruikt is ongedateerd. De dateringen in PiCarta en andere bibliografische bronnen spreken elkaar tegen. Hoe vind je de juiste datering? In dit geval: *Amsterdam bij nacht* van A.M. de Jong.

Volgens de journalist en schrijver A.M. de Jong (1888-1943) kende Amsterdam ooit een wild nachtleven. Hij schrijft hierover in een boekje getiteld *Amsterdam bij nacht*. Maar van wanneer dateert dat boekje? PiCarta kent drie exemplaren:

1920, circa 1920 en circa 1930 – dat is ons niet exact genoeg. Wat is het goede jaartal? PiCarta vermeldt als extra informatie dat het gaat om een herdruk uit *Het Leven*.

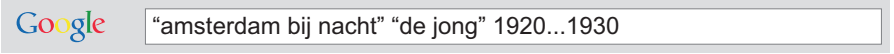
Van oudsher adverteren uitgevers in kranten. Dat maakt kranten ook belangrijke bronnen voor boekhistorisch onderzoek. Zoals gezegd is Historische Kranten de grootste, helderste en best doorzoekbare krantensite. Als je hier zoekt op "Amsterdam bij nacht" (tussen dubbele aanhalingstekens natuurlijk),

dan krijg je (momenteel) 163 resultaten. Die kun je sorteren door te klikken op het datumveld. De vroegste bron blijkt van 1855 te zijn, maar toen moest De Jong nog geboren worden. Dus kijken we meteen bij de eerste hit uit 1920, de *Nieuwe Rotterdamsche Courant* van 19 december 1920. Dit blijkt meteen raak, want hier vinden we een advertentie van *Het Leven* waarin een nieuwe artikelenreeks wordt aangekondigd:

Twee klikken verder – we zijn dan in februari 1921 aanbeland – zien we deze advertentie:

Hierin staat dat de reeks zojuist is afgerond. Er blijkt zoveel vraag naar oude afleveringen dat de uitgever heeft besloten ze te bundelen in een boekje dat niet in de handel zal komen, maar dat als premie aan nieuwe abonnees wordt gegeven. De daarop volgende advertenties spreken van 'Het telkens weer herdrukte succeswerk'.

Conclusie: *Amsterdam bij nacht* verscheen niet in 1920 of 1930, maar in 1921. Een laatste check op internet, die er bijvoorbeeld zo uit zou kunnen zien:



bevestigt dit, maar de informatie is nergens zo precies als in de Historische Kranten. Als je het zo stap voor stap opschrijft lijkt het wellicht alsof het zoeken lang heeft geduurd, maar het was binnen twee minuten gepiept, hooguit drie.

Over het zoeken van literatuur valt vanzelfsprekend nog veel meer te zeggen. Er zijn heel veel goede en gedegen catalogi. Welke je nodig hebt, kan afhangen van wat je zoekt, voor welke periode, in welke taal. Een ander goed beginpunt is WorldCat (www.worldcat.org) waar je momenteel kunt zoeken door 1,5 miljard items (o.a. boeken, artikelen, cd's' uit ruim tienduizend bibliotheken wereldwijd).

Casestudy 3

Onderzoek de geschiedenis van een woord. In dit geval: van het woord *boycot*.

Een goede eerste stap bij een begripsgeschiedenis is, zoals eerder gezegd: het *Woordenboek der Nederlandsche Taal* (WNT), het wetenschappelijke woordenboek van het Nederlands. Dit is een bron die je echt moet leren kennen, vooral door dingen te proberen. Als je in de bovenste zoekregel op **boycot** zoekt,

lijk je in eerste instantie slechts een kort artikel te vinden,

maar als je linksboven de twee vakjes aanvinkt wordt het artikel meteen een stuk langer:

De vroegste bewijsplaats in het WNT dateert van 1914. En het woord blijkt terug te gaan op de Ierse rentmeester Charles Boycott (met twee t's), die zelf werd geboycot.

Komt het woord ook nog op andere plaatsen in het WNT voor? Ja, die vind je bijvoorbeeld door te zoeken in het veld 'Woord in artikel':

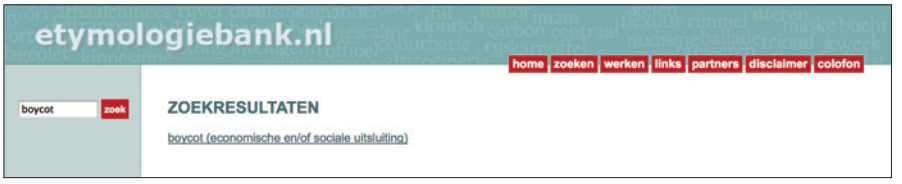
Door twee wildcards te gebruiken (ook eentje vóór het woord), vind je o.a. ook *geboycot*. Met deze zoekactie vind je nog enkele andere citaten, maar allemaal van ná 1914.

Volgens een ander woordenboek, de Grote Van Dale, kennen wij dit woord sinds 1910. De Engelsen kennen het echter al sinds 1880. Kwam het inderdaad pas dertig jaar later in het Nederlands terecht?

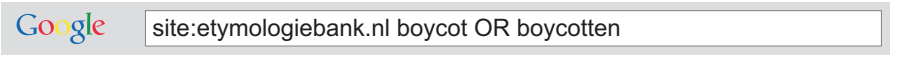
boycot

boy·cot
/bɔjkɔt/
zelfstandig naamwoord; de (m); meervoud: *boycots, boycotten*
(1910) Engels, naar *Charles Cunningham Boycott* (1832-1897), rentmeester in de Ierse County Mayo, die in 1880 als eerste werd geboycot

Tijd om over te stappen naar de tweede onmisbare bron voor woordgeschiedenissen: de Etymologiebank. Die kun je op twee manieren benaderen. Rechtstreeks (via de zoekregel op de website) levert 1 hit op:



Zoek met de site-operator via Google en voeg de meervoudsvorm toe



en je krijgt, in dezelfde site, *zeven* resulaten. We vinden hier van alles, maar de vroegste datering blijft 1910. Althans voor het zelfstandig naamwoord *boycot*. Het werkwoord *boycotten* blijkt volgens de Etymologiebank al in 1899 te zijn aangetroffen.

Is er een vroegere datering te vinden? Het wordt tijd voor onderzoek in contemporaine bronnen. We zoeken **boycot*** in de Historische Kranten (waar je helaas geen * kunt plaatsen vóór een woord, wel overal na de eerste letter). Tegelijkertijd beperken we via 'geavanceerd zoeken' de periode, we zoeken immers iets tussen 1880 en 1910.

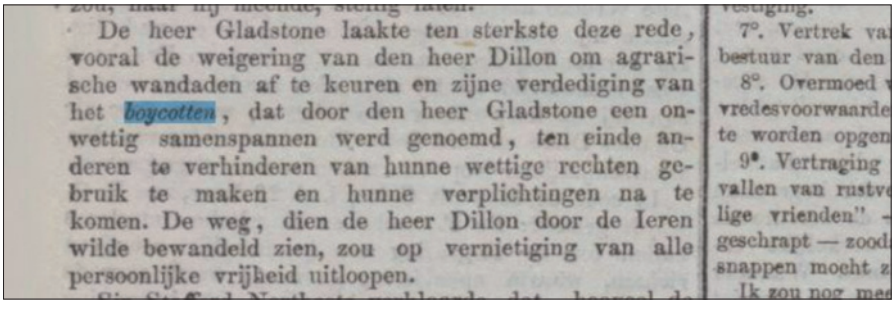
De eerste bron spreekt meteen uitvoerig over de kwestie, maar noemt Boycott alleen als eigennaam.

DE IERSCHE „LAND-LEAGUE".

De toestand in Ierland wordt nog dagelijks meer en meer gespannen: de *Land-league* treedt steeds dreigender op, en de reeks der agrarische misdrijven wordt voortdurend grooter. Zekere kapitein Boycott is eenige weken met zijn vrouw en dochter op zijn kasteel als het ware belegerd geweest, zonder dat iemand het

De Tijd, 17 november 1880.

De rel rond de Ierse rentmeester kreeg indertijd veel aandacht in de Nederlandse pers – alle details worden breed uitgemeten. In 1882 vinden we het werkwoord *boycotten* voor het eerst in een Nederlandse krant, maar wel cursief gezet en in een artikel over de Engelse en Ierse politiek, dus het is de vraag of dit als de vroegste Nederlandse bewijsplaats mag gelden (docent: 'Werk dit nader uit').



De heer Gladstone laakte ten sterkste deze rede, vooral de weigering van den heer Dillon om agrarische wandaden af te keuren en zijne verdediging van het *boycotten*, dat door den heer Gladstone een onwettig samenspannen werd genoemd, ten einde anderen te verhinderen van hunne wettige rechten gebruik te maken en hunne verplichtingen na te komen. De weg, dien de heer Dillon door de Ieren wilde bewandeld zien, zou op vernietiging van alle persoonlijke vrijheid uitloopen.

Algemeen Handelsblad, 28 mei 1882.

De volgende hit, uit 1884, heeft al niets meer met Engeland van doen:

oordeeling van het verraad
ɔdzakelijk geacht; terwijl
aucamps, van Brussel, de
bescherming wilden nemen.
id-minister van Financiën,
degene, die terstond na de
a Jacobs het harnas aantrok
staatrooverspolitiek nog
breken! Solvyns had als
steld: »de Senaat, de ver-
egeering gehoord hebbende,
ig hare afkeuring uit (*blâme*
over de hatelijke uitspat-

Brussel „Geboycotted".

Nu de liberale Brusselaars, getuige hun infame houding van Zondag, van meening zijn, dat zij alleen wat in de hoofdstad van het Rijk te zeggen hebben, ook als belangen op het spel staan, die niet slechts de stad, maar het geheele land raken, ontving het *Hand. v. Antw.* van verschillende zijden brieven, waarin de ernstige raad wordt gegeven, om Brussel te *boycotten*, m. a. w. er geen bestellingen meer te doen. En zelfs

Nieuwe Tilburgsche Courant, 14 september 1884.

Wanneer werd *geboycotted* (in de kop hierboven) tot *geboycot*? Wanneer duikt het zelfstandig naamwoord voor het eerst op? Wanneer verloren *boycott* en *geboycott* in het Nederlands hun tweede t? Werd het woord voornamelijk in politieke context gebruikt of ook daarbuiten? Was er een betekenisontwikkeling?

Je kunt het allemaal vinden door verder te klikken. Nog belangrijker: door het uit te zoeken en op te schrijven help je de wetenschap een stapje vooruit, want de dateringen in het WNT (1914), de Grote Van Dale (1910) en de Etymologiebank (1899) blijken achterhaald. Zonder twijfel kunnen ook de dateringen in Franse en Duitse woordenboeken worden aangescherpt. Voor je het weet ben je aan het uitzoeken hoe snel en op welke manier dit nieuwe begrip zich over Europa heeft verbreid (razendsnel). En hoe heette maatschappelijke uitsluiting eigenlijk voordat we dit internationaal *boycotten* gingen noemen?

Overigens is het niet altijd eenvoudig om iets in de Historische Kranten te vinden. In principe worden hits gemarkeerd (gehighlight), maar helaas niet altijd. En soms zijn ze nauwelijks te vinden op een krantenpagina. Oplossing één: klik hier

op het icoontje *tekst* en zoek vervolgens met Ctrl + f(ind) naar het gezochte woord. Je kijkt nu in de tekstlaag achter de afbeelding van de pagina. Hier kun je meteen zien hoeveel ocr-fouten er in oude teksten worden gemaakt.

Oplossing twee: klik op het buitenste icoontje, *pdf*. Daarmee download je de krant. Nu kun je via het zoekvenster van bijvoorbeeld Adobe Reader naar het trefwoord zoeken.

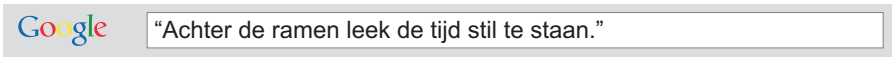
Casestudy 4

In een literaire tekst staat een citaat zonder bronvermelding. De opdracht luidt: zoek de auteur van dit citaat. In dit geval: een citaat bij Karel van het Reve.

In 1970 schreef Karel van het Reve, een schrijver die graag met citaten strooide, in *Marius wil niet in Joegoslavië wonen* (p. 21-22): 'Achter de ramen leek de tijd stil te staan'. Welke lezer weet van welk boek dit de laatste regel is?

De bezorgers van het *Verzameld werk* van Karel van het Reve wilden deze vraag niet onbeantwoord laten. Hoe kun je, met een beetje geluk, relatief snel de bron voor dit citaat vinden?

Logische eerste stap, het hele citaat tussen dubbele aanhalingstekens in Google:



Helaas, aantal resultaten: 0.

Hoe nu verder? Wat nog wel eens helpt is dit: isoleer de meest kenmerkende woorden uit het citaat en probeer het opnieuw, in twee fragmenten, zodat er andere dingen tussen kunnen staan. Dus:

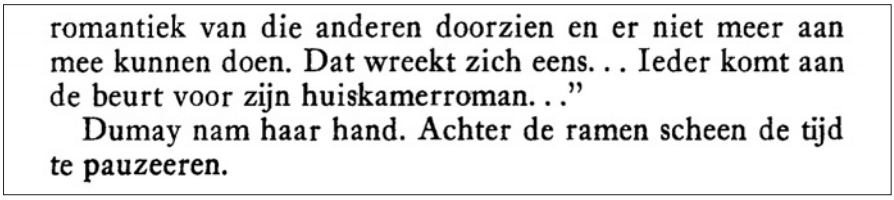
Dit levert volgens Google 422 resultaten op, te veel om snel te bekijken.

Dan maar zoeken met sterretjes, en binnen de DBNL, want die site bevat momenteel de grootste collectie Nederlandse taal- en letterkunde. Eerste poging:

Nu zijn er nog maar zeven resultaten over, die allemaal naar Menno ter Braak verwijzen. Hier de eerste twee:

Er staat echter niet: 'Achter de ramen **leek** de tijd **stil** te **staan**', maar 'Achter de ramen **scheen** de tijd te **pauseeren/pauzeren**' – daarom was het letterlijke citaat dus niet te vinden.

Is het wel de laatste regel van een boek? Ja, namelijk van Menno ter Braaks roman *Dr. Dumay verliest*... (de puntjes horen bij de titel) uit 1933. Bij de DBNL is dit boek in twee formaten aanwezig, in xml en als pdf. Wie voor de zekerheid de pdf opent, ziet als slotzinnen op de laatste pagina:

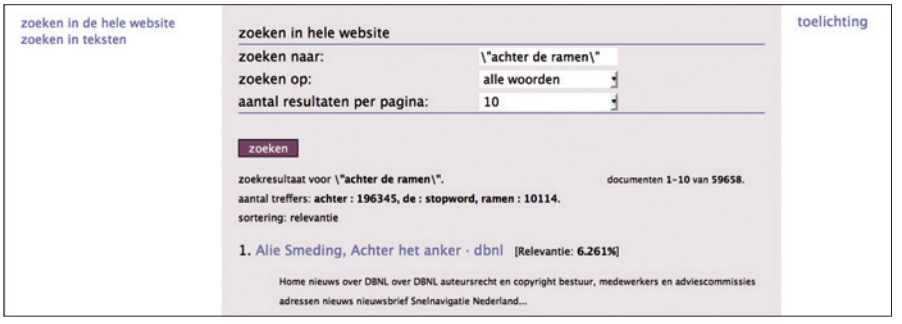

romantiek van die anderen doorzien en er niet meer aan mee kunnen doen. Dat wreekt zich eens. . . Ieder komt aan de beurt voor zijn huiskamerroman. . ."

Dumay nam haar hand. Achter de ramen scheen de tijd te pauzeeren.

Conclusie: Karel van het Reve citeerde uit het hoofd en vergiste zich bij een paar woorden.

Tot slot: was dit citaat te vinden geweest als je niet via Google maar *rechtstreeks* bij de DBNL was gaan zoeken? We proberen het met "achter de ramen", tussen dubbele aanhalingstekens, want volgens de toelichting kan dat:

Dit levert echter ruim 59.000 resultaten op, waarom is niet precies duidelijk. Kennelijk gaat er van alles mis, want in alle hoogste resultaten komt alleen het woord *achter* voor. En kennelijk wordt het lidwoord *de*, zelfs bij een zoekopdracht tussen aanhalingstekens, als *stopword* gezien:



Hoe het ook zij, in dit geval was het duidelijk slimmer om niet rechtstreeks bij de DBNL te zoeken, maar via Google met behulp van de zoekoperator **site:** (site:dbnl.org). Overigens kun je bij de DBNL nog op allerlei andere manieren zoeken. Met een kleine test ben je er zo achter wat het best werkt voor jouw zoekvraag: rechtstreeks of via Google.

Denk niet te snel: voor mijn studie zijn deze voorbeelden niet van toepassing. Het gaat hier om de zoektrucs, niet speciaal om de content. Uitdaging: welke combinatie van zoekoperatoren kan ik het best inzetten om snel de informatie te vinden die wél precies bij mijn studie of onderzoek past?

Casestudy 5

Van de bekende uitdrukking *dat slaat als een tang op een varken* bestaan in het hedendaagse Nederlands allerlei varianten. Hoe kun je die vinden? Welke komen het vaakst voor? En waar worden ze het meest gebruikt: in Nederland of in Vlaanderen?

Eerste voor de hand liggende zoekactie:

Dit levert meteen twee hits op over de herkomst van de uitdrukking, met als belangrijkste het spreekwoordenboek van F.A. Stoett, verschenen tussen 1923 en 1925.

Stoett geeft informatie over de herkomst van de uitdrukking en allerlei interessante historische varianten, zoals *dat sluyt ghelyck ses vingeren in een handtschoen*, maar we waren op zoek naar moderne varianten. Hoe vind je die?

Door creatief gebruik te maken van de tekens * en -.

Je zoekt *dat slaat als een tang op een varken*, maar dan zonder *tang* en zonder *varken*. Dus:

Al bij de eerste tien resultaten vind je:

- slaat als een lul op een drumstel
- slaat als een knul op een drumstel
- slaat als een drol op een slagroomtaart
- slaat als een slok op een borrel

Op basis van deze resultaten kun je verder gaan schrappen:

“slaat als een * op een *” -tang -varken -lul -drumstel -knul -drol -slagroomtaart -slok -borrel

Er volgt nu van alles: *slaat als een tiet op een stoeprand, slaat als een schuiftrompet op een houtvlot, slaat als een hamer op een banaan*, enzovoorts. Ook die kun je, met behulp van het minteken, allemaal wegfilteren. Gevolg: je resultatenlijst wordt steeds korter, terwijl je nieuwe varianten blijft vinden.

Wat je nu niet vindt, is: *slaat als **’n** drol op **’n** slagroomtaart*, je zocht immers op “slaat als **een** * op **een** *”. Maar zou je dat zelf zo schrijven? In zo’n informele uitdrukking? Beter was het dus geweest om nog een extra sterretje te plaatsen, dus: “slaat als * * op * *” -tang -varken.

Uiteindelijk maak je van je zoekopdracht een lange ketting, die maximaal 128 woorden mag bevatten. Dit is een opdracht waar je geduld voor nodig hebt, want je hebt ook nog varianten als *dat slaat als kut op Dirk* en *dat **rijmt** als x op y*.

Hoe vaak komt een bepaalde variant nu voor? Dat onderzoek je door gericht naar die ene variant te zoeken, bijvoorbeeld zo: "slaat als * drol op * slagroomtaart". Als je wilt vaststellen of een bepaalde variant wordt gebruikt in Nederland en/of België, voeg je aan je zoekopdracht simpelweg de site-operator met een landcode toe. Dus:

site:nl "slaat als * drol op * slagroomtaart" en site:be "slaat als * drol op * slagroomtaart"

Met de aantallen die Google vervolgens opgeeft, moet je overigens oppassen. Voorbeeld:

Volgens Google komt de uitdrukking *slaat als een tang op een varken* (waarbij *een* ook gespeld kan zijn als *'n*) op Belgische websites zo'n 10.500 keer voor. Maar is dat ook zo? Er is helaas maar één, vervelende, manier om daar achter te komen, en dat is stapsgewijs doorklikken naar de laatste resultatenlijst:

In dit geval blijken er, anders dan je zou verwachten, slechts 10 resultatenlijsten te zijn:

en aangezien iedere pagina – in mijn geval – 10 resultaten telt, is de werkelijke frequentie 100, in plaats van 'ongeveer 10.500' (als je Google Instant uitschakelt, kun je 100 resultaten per pagina krijgen). Zelfs 100 blijft een grove telling, want sommige webpagina's staan meer dan één keer tussen de resultaten en het kan ook zijn dat de gezochte uitdrukking vaker dan één keer op een pagina voorkomt.

Kortom, pas op met de aantallen die Google noemt – je kunt ze zeker niet klakkeloos overnemen.

Alles bij elkaar is dit dus geen opdracht waar je binnen drie minuten mee klaar bent, maar het is wel mooi om te zien hoe creatief er wordt gevarieerd op een uitdrukking die al sinds de zestiende eeuw in het Nederlandse taalgebied voorkomt. (Docent: zoek varianten op *domweg gelukkig in de Dapperstraat* en de volgende klassieke dichtregels; vind een betrouwbare bron voor het origineel; analyseer de varianten en leg stap voor stap uit hoe je… enzovoorts.)

Ngrams

Een interessant nieuw instrument voor frequentietellingen is de Ngram Viewer van Google (**www.ngrams.googlelabs.com**). Hier kun je een selectie van ruim vijf miljoen boeken uit Google Books doorzoeken op reeksen van maximaal vijf woorden ('n-grammen', zoals Google ze noemt, waarbij *n* staat voor een reeks van maximaal vijf). De resultaten worden gepresenteerd in een grafiek. De Viewer werkt helaas nog niet voor het Nederlands, maar wel voor onder andere het Engels (met een onderscheid tussen Brits en Amerikaans), Duits, Frans, Spaans en Russisch. De meeste boeken in Google Books zijn overigens gelezen met ocr, met alle fouten van dien.

Casestudy 6

Je krijgt een vage verwijzing naar een bepaald oud tijdschrift ('deel 2 van *De Gids*, waarschijnlijk in 1850 of 1851'). Er zit niks anders op dan van die jaargangen de inhoudsopgaven door te nemen. Hoe pak je dat aan?

De meeste mensen besteden niet veel aandacht aan url's (webadressen), zeker niet als die er ingewikkeld uitzien. Toch kun je er veel aan hebben. Vooral bij grote en professionele websites hebben url's een vaste structuur, die meestal verwijst naar een mappenstructuur en/of een achterliggende database. Daar kun je je voordeel mee doen.

In dit geval moet je een vage verwijzing nazoeken: 'deel 2 van *De Gids*, waarschijnlijk in 1850 of 1851'. De DBNL is alle jaargangen van *De Gids*, ooit een zeer invloedrijk maandblad, aan het digitaliseren en heeft al veel gedaan. Veel mensen zullen hier beginnen bij de DBNL:

om vervolgens te gaan klikken.

Klik 1:

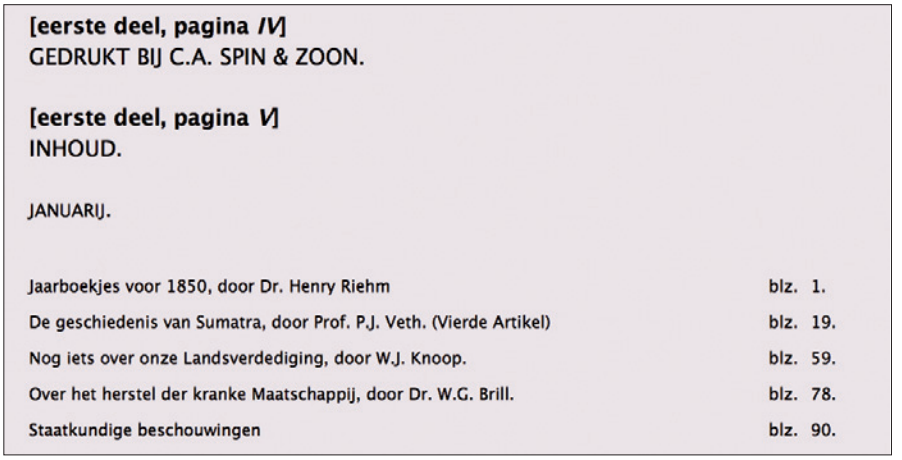
Klik 2:

Klik 3:

Je denkt dat je nu bent waar je wezen wilt, want je zit immers in de inhoudsopgave. Maar in de praktijk gaat het hier om links naar de inhoud van de xml-versie van het tijdschrift, wat niet hetzelfde is. Je kunt hier bijvoorbeeld niet zien waar deel één eindigt en deel twee begint (of in welke maand iets is verschenen, wat ook erg handig kan zijn).

De echte inhoudsopgave blijkt te vinden onder het kopje *verantwoording*:

Nu zie je (klik 4) per maand de inhoudsopgave



Helaas, hier vond je niet wat je zocht, dus op naar jaargang 1851. Hier zijn, je begrijpt het al, weer veel klikken voor nodig (om precies te zijn: zeven).

Kan dat sneller? Ja, door naar de url te kijken.

De url van de inhoudsopgave van jaargang 1850 van *De Gids* ziet er zo uit:

http://www.dbnl.org/tekst/_gid001185001_01/colofon.php

De url van jaargang 1851 van *De Gids* ziet er zo uit:

http://www.dbnl.org/tekst/_gid001185101_01/colofon.php

Wat is het verschil? Inderdaad: gid001**1850**01 en gid001**1851**01.

Wil je nu even snel de inhoudsopgave van 1853 zien? Dat doe je dus door één cijfer in de url te veranderen, wat een stuk sneller gaat dan zeven keer klikken.

Wil je bij de DBNL snel tien, twintig of meer jaargangen van een tijdschrift downloaden? Dat kan met dezelfde truc. Uitgangspunt is dan bijvoorbeeld deze url:

http://www.dbnl.org/tekst/_gid001186001_01/downloads.php

Voor de goede orde: soms is in een reeks een bepaalde jaargang (nog) niet gedigitaliseerd. Dan krijg je een foutmelding. Door op de gewone manier te zoeken (terug naar de oorspronkelijke lijst met titels bijvoorbeeld) zie je al snel wat er aan de hand is. Maar meestal loopt het allemaal soepel en scheelt gebruikmaken van de url-structuur je veel kliktijd.

Zoals gezegd is deze truc bij talloze websites toe te passen. Nog één voorbeeld, tot slot: het *Utrechts(ch) Nieuwsblad*, dat van 1893 tot 1967 te raadplegen is via de site van het Utrechts Archief (www.hetutrechtsarchief.nl). De gewone toegang tot deze bron, via de website van het archief, is uiterst moeizaam, dus het is een opluchting als je hebt ontdekt hoe de url-structuur in elkaar steekt (zie het kader 'Een url-structuur vinden'). Zo ziet de url van het *Utrechtsch Nieuwsblad* van 22 juni 1942 eruit:

http://www.hetutrechtsarchief.nl/collectie/kranten/un/1942/0622

Kortom: 22 juni 1942 is *1942/0622* geworden, en *Utrechtsch Nieuwsblad* /un/. Doordat je nu niet langer wordt beperkt door de zoekmogelijkheden van het archief, kun je in deze interessante bron, die niet is opgenomen in de Historische Kranten, veel meer vinden. Probeer eens een paar dingen uit, bijvoorbeeld:

site:hetutrechtsarchief.nl/collectie/kranten/un **koperdief OR koperdieven**
of:
site:hetutrechtsarchief.nl/collectie/kranten/un **diefstal 1893...1900**
dan wel:
site:hetutrechtsarchief.nl/collectie/kranten/un/**1963/1123** kennedy
(om snel te vinden wat deze krant schreef over de moord op Kennedy.)

Op deze manier kun je makkelijk een necrologie van een schrijver vinden, een bepaald jubileum (site:hetutrechtsarchief.nl/collectie/kranten/un jubileum "koningin Juliana") enzovoorts.

Een url-structuur vinden

Hoe achterhaal je de url-structuur van bijvoorbeeld *Elsevier Maandschrift*?

1. Zoek:

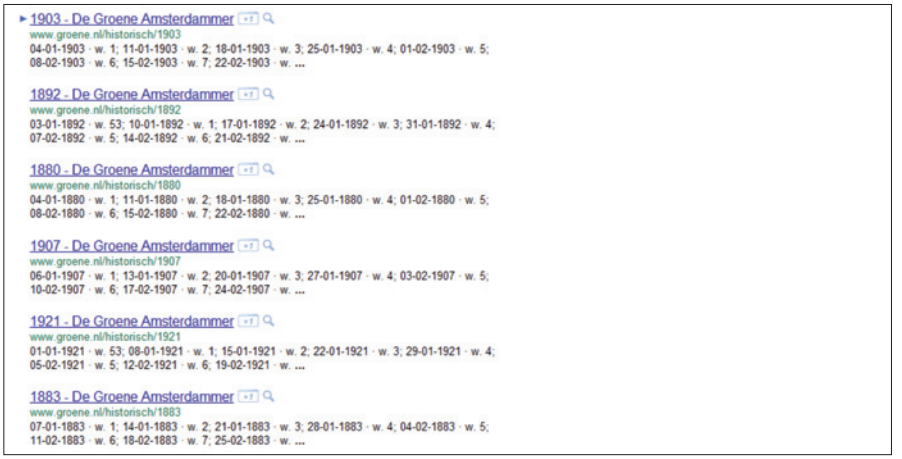
2. Eerste resultaat:

3. Zoek:

Niet schrikken, je krijgt honderden resultaten, maar vergelijk er twee en je ziet hoe de url-structuur is opgebouwd.

www.elseviermaandschrift.nl/EGM/1909/07/19090701/EGM.../story.pdf
www.elseviermaandschrift.nl/EGM/1929/07/19290701/EGM.../story.pdf

Plezier in? Probeer dan ook eens:



Niet alle databanken laten dit overigens toe, maar veel wel.

Casestudy 7

Zoek binnen een database een woord of naam met behulp van wildcards. Houd daarbij rekening met twee dingen: spellingvariatie en ocr-fouten. Hoe pak je dit aan? In dit geval: zoek binnen Early Dutch Books Online naar het woord *komkommertijd*.

Je zou willen dat het Nederlands van oudsher één spelling had gekend (liefst een heldere en makkelijke), dat geen schrijver of zetter ooit een spelfout had gemaakt en dat optische tekenherkenning (ocr) honderd procent foutloos was. Dat alles zou het werk van veel onderzoekers een stuk makkelijker maken.

In feite bestaat er pas sinds 1804 een officiële spelling, die meermaals is gewijzigd en waar volop van is afgeweken. Op dergelijke spellingveranderingen kun je bij het zoeken trouwens nog wel anticiperen, omdat er een bepaalde lijn in zit. Dat geldt helaas niet voor de talloze ocr-fouten in ouder drukwerk.

We staan dus voor een uitdaging, en daarbij kunnen twee kleine tekens ons grote diensten bewijzen: ? (ter vervanging van één willekeurig teken) en * (ter vervanging van een willekeurig aantal tekens).

Hoe ga je daar in de praktijk mee aan de slag? We zoeken naar het woord *komkommertijd* in Early Dutch Books Online (EDBO), zoals eerder gezegd een collectie van 11.240 boeken uit de periode 1781-1800. In dit bestand van 600 miljoen woorden staan veel ocr-fouten, en *tijd* werd toen ook gespeld als *tyd* (een goede bron voor spellingvarianten: het WNT, en natuurlijk ook oude spellinggidsen, maar die staan nog niet allemaal online).

Een korte inhoudelijke toevoeging: tot voor kort nam men aan dat *komkommertijd* in 1871 voor het eerst was aangetroffen. Dat maakte het onwaarschijnlijk dat het om een leenvertaling van het Duitse *Sauregurkenzeit* ('zurebommentijd') ging, want dit was al sinds het eind van de 18de eeuw bekend. Is in EDBO een oudere bron voor *komkommertijd* te vinden?

Ja, door simpelweg op *komkommertijd* te zoeken, want in die schrijfwijze komt het woord één keer in EDBO voor, in een bron uit 1792. Met behulp van wildcards vind je nog drie andere bronnen.

Zoeken met wildcards is een kwestie van nadenken en uitproberen. *Tijd* telt vier letters, want de i en de j tellen voor twee. *Tyd* telt drie letters. De zoekopdracht t??d (vier tekens), zal dus alleen *tijd* opleveren, niet *tyd*. Beter is het dus om naar t*d te zoeken, want met een sterretje vervang je immers een *willekeurig* aantal letters: één of veel meer.

Opmerkelijk ervaringsfeit: weinig alfa's schrikken van eenvoudige wiskundige formules als *7 cm x 5 cm : 2 = 17,5 cm²* (op school gehad), maar als ze zoekopdrachten zien als: *komk????t*d* **of** *site:dbnl.org "achter de ramen * *"* **dan haken ze al snel af ('dat is mij echt te ingewikkeld'). En dat terwijl je, vaak binnen seconden, veel baat kunt hebben bij zo'n geavanceerde zoekopdracht.**

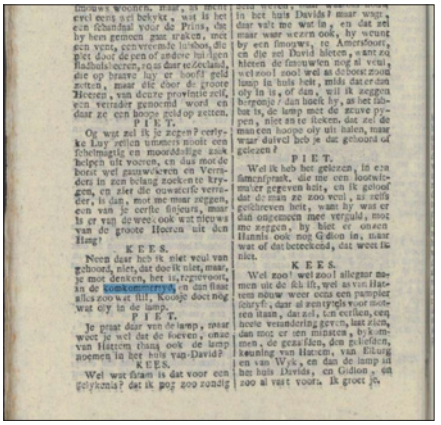
Komkommer telt negen tekens. Een logische zoekopdracht is dus: **kom??????t*d**.
Dit levert echter alleen dezelfde bron uit 1792 op. Hetzelfde geldt voor de zoekopdrachten:

komk?????t*d
komko????t*d
komkom???t*d

Dan maar het minder beredeneerde: **k*mk*mm*d***
Dit blijkt de beste opdracht, want hiermee vang je (binnen 24 zoekresultaten) vier vliegen in één klap:

1 x de ocr-fout *kimkommerryd* uit 1786-1787,
1 x de ocr-fout *komkommerryd* uit 1787,
1 x *komkommertijd* uit 1792, en
1 x *komkommertjens-tyd* uit 1800.

De Haagsche correspondent, 1786-1787. In de ocr-laag staat *kimkommerryd*.



Het Saturdags Kroegpraatje van 14 juli 1787. In de ocr-laag staat *komkommerryd*.

Wat het zoeken betreft volgt hieruit: staar je niet blind op het aantal tekens van het woord dat je zoekt, in dit geval: de 13 tekens van *komkommer**tijd*** en de 12 tekens van *komkommer**tyd***. Als iemand een koppelteken had gebruikt (*komkommer-tijd*), dan hadden deze aantallen ook al niet geklopt. Door sterretjes te gebruiken kun je wel meer ruis krijgen (*komkommervelden, komkommerzaad, komkommersaladen*), maar meestal klik je hier snel doorheen.

Ook inhoudelijk is er een conclusie: nu is vastgesteld dat het Nederlandse *komkommertijd* niet pas in 1871 in het Nederlands opduikt, maar kort nadat de Duitsers begonnen te spreken van *Sauregurkenzeit,* is het zeer waarschijnlijk dat *komkommertijd* een leenvertaling uit het Duits is.

Overigens is deze hele exercitie volgens de handleiding bij de eerste release van EDBO helemaal niet mogelijk, want bij 'veelgestelde vragen' staat:

- **Als ik op een woord als "opstellen" zoek, wil ik ook bladzijden met "opftellen" kunnen vinden, omdat ik weet dat een "lange s" door de OCR- software vaak als "f" gelezen wordt. Waarom kan dit niet?**
  Dit komt omdat fuzzy search op dit moment niet mogelijk is. U dient op beide varianten apart te zoeken.
- **Als ik op een woord als "herder" zoek, wil ik ook resultaten met "herders" zien. Waarom kan dit niet?**
  Dit komt omdat fuzzy search op dit moment niet mogelijk is. U dient op beide varianten apart te zoeken.

Onzin dus, want *opftellen* en *opstellen* vind je in één keer met de zoekopdracht **op?tellen**, en *herder* en *herders* door naar **herder OR herder?** te zoeken (want toevoeging van de operator OR blijkt ook te kunnen). In de tweede release van EDBO zal dit worden rechtgezet.

Les: trek je niet te veel aan van handleidingen en toelichtingen. Probeer standaard of wildcards en booleaanse operatoren wel of niet werken. Overigens werken wildcards slechts zelden aan het begin van een woord (**?oord** of ***oord** geeft bij EDBO en de Historische Kranten geen resultaat, maar bij het WNT wel). Hoe groter de database, hoe kleiner de kans dat *trunceren* aan de voorzijde (zoals dit officieel heet) lukt, want dit kost veel geheugenruimte.

Exalead

Momenteel is er slechts één zoekmachine op internet waarmee je – niet vóór het woord maar na de eerste letter – een of meer asterisken kunt plaatsen (zonder spatie). Zie: **www.exalead.com/search**

Exalead heeft weer andere interessante zoekoperatoren, zoals **soundslike:** (fonetisch), **spellslike:** (zoekt meteen spellingvarianten), **before:** en **after:** (zoekt vóór en na een bepaalde datum). Voor een helder overzichtslijstje, zie **www.exalead.com/search/web/search-syntax**.

Casestudy 8

Huilende mannen en huilende vrouwen – werden ze anders beschreven in bronnen uit een bepaalde periode? In dit geval: de *Vaderlandsche Letteroefeningen* tussen 1800 en 1810.

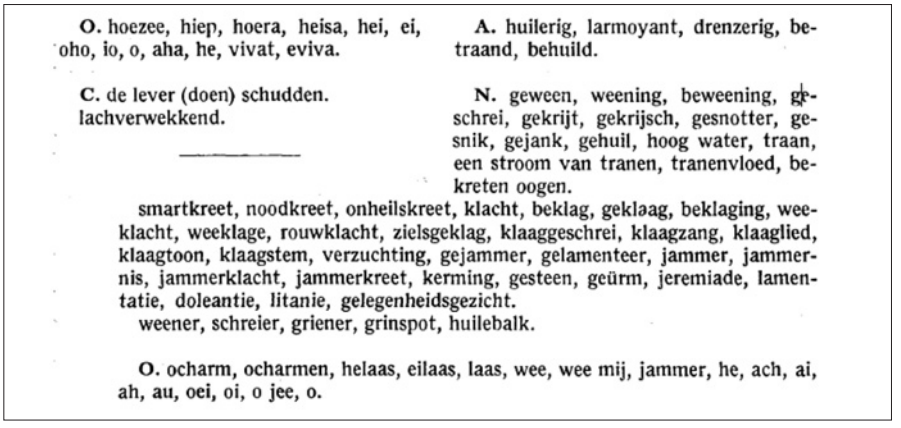
Er zijn met gemak meer interessante opdrachten te verzinnen die kunnen worden beantwoord met de digitale bronnen die ons nu al ter beschikking staan. Bijvoorbeeld: onderzoek hoe er in kranten werd gereageerd op de dood van een bekende schrijver. Of: ga na hoe er werd gereageerd op de introductie van de radio, roltrap, lift, telefoon of bioscoop. Dan wel: op de introductie van de tomaat, sinaasappel of kiwi (zoek niet alleen naar krantenartikelen maar ook naar beeldmateriaal en liedjes). Nog een: je hoort vaak dat het taalgebruik in de Tweede Kamer steeds grover en informeler wordt. Maar is dat ook zo? Ga dit na aan de hand van enkele woorden en uitdrukkingen die bekendstaan als grof en informeel; vergeet vooral *flapdrol* niet.

In principe is het materiaal voor de antwoorden op deze vragen te vinden met behulp van zoektechnieken die hierboven al zijn beschreven. Dat geldt niet, althans niet helemaal, voor een type onderzoek dat op dit moment veel aandacht krijgt in de geesteswetenschappen, namelijk onderzoek met behulp van een woord- of woordenwolk.

Het idee is eenvoudig: je wilt bijvoorbeeld onderzoek doen naar emoties in teksten. In dit geval: naar huilende mensen (krokodillen huilen ook, maar zij schreien hun krokodillentranen hoogstwaarschijnlijk om overtollig zout af te scheiden). Mensen huilen uit blijdschap en verdriet, maar tekstueel maakt dat niet veel uit. Als ergens over huilende mannen, vrouwen of kinderen wordt gesproken, zullen daar woorden in voorkomen als *huilen*, *schreien*, *wenen*, *grienen*, *tranen*, *biggelen*, *blèren*, *snikken*, *jammeren*, enzovoorts.

Eerste taak is dus om de woorden en uitdrukkingen die met huilen zijn geassocieerd bij elkaar te zetten. Zo'n groep wordt, overigens pas sinds kort, een *woordenwolk* genoemd.

Zijn er hulpmiddelen om zo'n wolk samen te stellen? Ja, want in het verleden zijn er diverse naslagwerken verschenen waarin dit soort betekenisverwante woorden en uitdrukkingen bij elkaar zijn gezet. Een bekend werk is *Het juiste woord* van de Vlaamse priester L. Brouwers (een snelle blik in PiCarta leert dat er van zijn werk tussen 1931 en 2002 zes edities zijn verschenen). In de kolom *vreugde – droefheid* geeft Brouwers onder meer deze opsomming:



O. hoezee, hiep, hoera, heisa, hei, ei, oho, io, o, aha, he, vivat, eviva.

A. huilerig, larmoyant, drenzerig, betraand, behuild.

C. de lever (doen) schudden. lachverwekkend.

N. geween, weening, beweening, geschrei, gekrijt, gekrijsch, gesnotter, gesnik, gejank, gehuil, hoog water, traan, een stroom van tranen, tranenvloed, bekreten oogen.

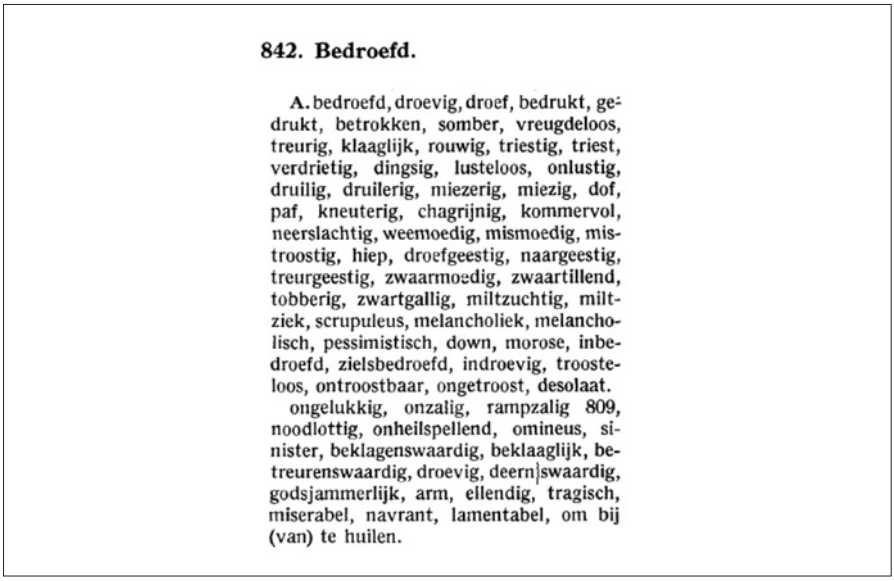
smartkreet, noodkreet, onheilskreet, klacht, beklag, geklaag, beklaging, weeklacht, weeklage, rouwklacht, zielsgeklag, klaaggeschrei, klaagzang, klaaglied, klaagtoon, klaagstem, verzuchting, gejammer, gelamenteer, jammer, jammernis, jammerklacht, jammerkreet, kerming, gesteen, geürm, jeremiade, lamentatie, doleantie, litanie, gelegenheidsgezicht.

weener, schreier, griener, grinspot, huilebalk.

O. ocharm, ocharmen, helaas, eilaas, laas, wee, wee mij, jammer, he, ach, ai, ah, au, oei, oi, o jee, o.

Brouwers, *Het juiste woord* (1931).

En bij *bedroefd*:



842. Bedroefd.

A. bedroefd, droevig, droef, bedrukt, gedrukt, betrokken, somber, vreugdeloos, treurig, klaaglijk, rouwig, triestig, triest, verdrietig, dingsig, lusteloos, onlustig, druilig, druilerig, miezerig, miezig, dof, paf, kneuterig, chagrijnig, kommervol, neerslachtig, weemoedig, mismoedig, mistroostig, hiep, droefgeestig, naargeestig, treurgeestig, zwaarmoedig, zwaartillend, tobberig, zwartgallig, miltzuchtig, miltziek, scrupuleus, melancholiek, melancholisch, pessimistisch, down, morose, inbedroefd, zielsbedroefd, indroevig, troosteloos, ontroostbaar, ongetroost, desolaat.

ongelukkig, onzalig, rampzalig 809, noodlottig, onheilspellend, omineus, sinister, beklagenswaardig, beklaaglijk, betreurenswaardig, droevig, deern]swaardig, godsjammerlijk, arm, ellendig, tragisch, miserabel, navrant, lamentabel, om bij (van) te huilen.

Oude synoniemenverzamelingen kunnen dus nuttige hulpbronnen zijn bij het samenstellen van een woordenwolk, maar het blijft een lastige klus – zeker naarmate je verder teruggaat in de tijd. Meestal zul je er gedurende je onderzoek achter komen welke woorden je zeker niet mag missen. Voor een onderzoek naar joden zul je in het Nederlandse taalgebied bijvoorbeeld ook moeten zoeken naar: *Hebreeër, Israeliet, jehoede, jid, Judeeër, Moos, pakslinger, parg, piechem, poerem, Polakkie, smous* en *snaggel* (de opsomming is niet compleet). Tel daar alle spellingvarianten bij op (*schnaggel, smaus, pigem*), en je ziet wat een klus het gaat worden om dit allemaal na te zoeken in bestanden van honderden miljoenen woorden vol onvoorspelbare ocr-fouten.

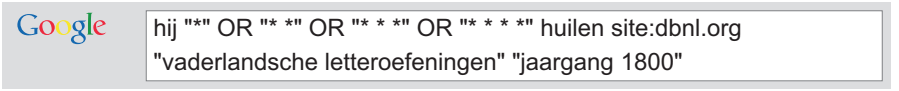
In deze casestudy, die wellicht meer voor gevorderden dan voor beginners geschikt is, wil je ook nog zoeken naar een verband tussen huilen en geslacht (man, vrouw). Dit kun je doen door Google te dwingen om een zogenoemde *proximity search* uit te voeren, waarbij twee (of meer) woorden op een bepaalde afstand van elkaar mogen staan. Dat doe je bijvoorbeeld zo:

Vertaald in gewoon Nederlands: ik zoek *hij* gevolgd door *huilen*. Tussen deze woorden mogen één, twee, drie of vier willekeurige woorden staan.

Google geeft onder andere als resultaten:

- als **hij** gaat **huilen** om aandacht (1 woord tussen *hij* en *huilen*)
- Als andere mensen hem benaderden verstijfde **hij** meteen, ging **huilen** (2 woorden tussen *hij* en *huilen*)
- **hij** heeft voornamelijk zitten **huilen** (3 woorden tussen *hij* en *huilen*)
- kan **hij** gefrustreerd raken en gaan **huilen** (4 woorden tussen *hij* en *huilen*)

Nu we weten dat deze truc werkt, kunnen we 'm toepassen binnen de DBNL, waar de meest foutloze collectie oude Nederlandse teksten te vinden is. Bijvoorbeeld met deze zoekopdracht:



Dus: ik zoek *hij* gevolgd door *huilen*. Tussen deze woorden mogen één, twee, drie of vier willekeurige woorden staan. Zoek dit op de *site* van de DBNL, in het tijdschrift *Vaderlandsche Letteroefeningen*, te beginnen in *jaargang 1800*. Dit levert meteen een interessant resultaat op, namelijk:

Je zou de opdracht nog kunnen uitbreiden: niet alleen jaargang 1800, maar meteen ook 1801 (OR "jaargang 1801"), 1802 (OR "jaargang 1802"). *Hij* kan worden vervangen door *hy*, *zij* of *zy*, *huilen* door *wenen*, *weenen*, *schreien*, *jammeren*, in plaats van maximaal vier willekeurige woorden kun je er tien tussen zetten – je kunt het net zo ingewikkeld maken als je zelf wilt.

Een onderzoek als dit kost tijd, maar die valt in het niet bij de tijd die tot voor kort vereist was om zelfs maar een fractie van het nu doorzochte materiaal bij elkaar te krijgen, laat staan te doorzoeken.

Samenvatting van hoofdstuk 4 in een woordenwolk

5 THE NEXT LEVEL

Wie al het voorgaande heeft gelezen, zal wellicht denken: kan dat allemaal niet wat eenvoudiger? Al dat gepiel met sterretjes, plussen, minnetjes, vraagtekens, url-aanpassingen en operatoren.

Er zijn twee tools die veel gemak kunnen opleveren: *webcrawlers* (ook wel *crawlers*, *webspiders* of *spiders* genoemd) en indexeringssoftware.

Webcrawlers zijn robots die websites indexeren: zij zetten alle woorden die op een site voorkomen in een lijst. De zoekmachine van Google is in feite een webcrawler – een heel grote en buitengewoon geavanceerde, maar uiteindelijk toch een robot die internet afgraast en alle gevonden informatie in een eigen database plaatst. Daarin zoek je als gebruiker.

Met een webcrawler voor thuisgebruik kun je het bescheidener aanpakken. Tegen zo'n robot kun je bijvoorbeeld zeggen: hier heb je een lijst van tien of twintig websites die ik doorlopend raadpleeg. Ik wil dat je ze dagelijks / wekelijks / maandelijks bezoekt en alles wat je er tegenkomt opslaat in een index. Meestal zal de crawler een kopie van de webpagina's op je eigen computer zetten. Wat overigens, zelfs bij grote websites, minder ruimte inneemt dan menigeen denkt.

Indexeringstools werken op dezelfde manier, maar dan voor documenten op je computer: mail, Worddocumenten, pdf's, de meeste indexeringstools kunnen honderden formaten aan.

Veel tools combineren beide functies: ze kunnen zowel websites als bestanden op je eigen computer indexeren. Er zijn gratis indexeringstools, maar de pakketten waarvoor je moet betalen zijn beter en geavanceerder (je hebt al goede crawlers en indexeringstools voor een paar tientjes).

Waarom zou je hierin investeren? Omdat je met behulp van webcrawlers en indexeringstools nog een stuk slimmer kunt zoeken dan met de huidige zoekmachines op internet. Je kunt er van alles en nog wat mee, maar zelf gebruik ik vooral deze functies:

1. de *proximity search*
   Bijvoorbeeld: geef mij A maximaal 5 / 10 / 20 posities van B; geef mij A in dezelfde alinea als B; geef mij A voor B of juist andersom; enzovoorts. Dit scheelt, in vergelijking met Google, een hoop gedoe met asterisken.

2. wildcards (? en *) aan de voorzijde van een of meer woorden. Voorbeelden:
   - met *heid vind je alle woorden die hierop eindigen;
   - met *fiets* vind je niet alleen *bakfiets* en *fietstas*, maar ook aan*fiets*en, brom*fiets*helm, enzovoorts.

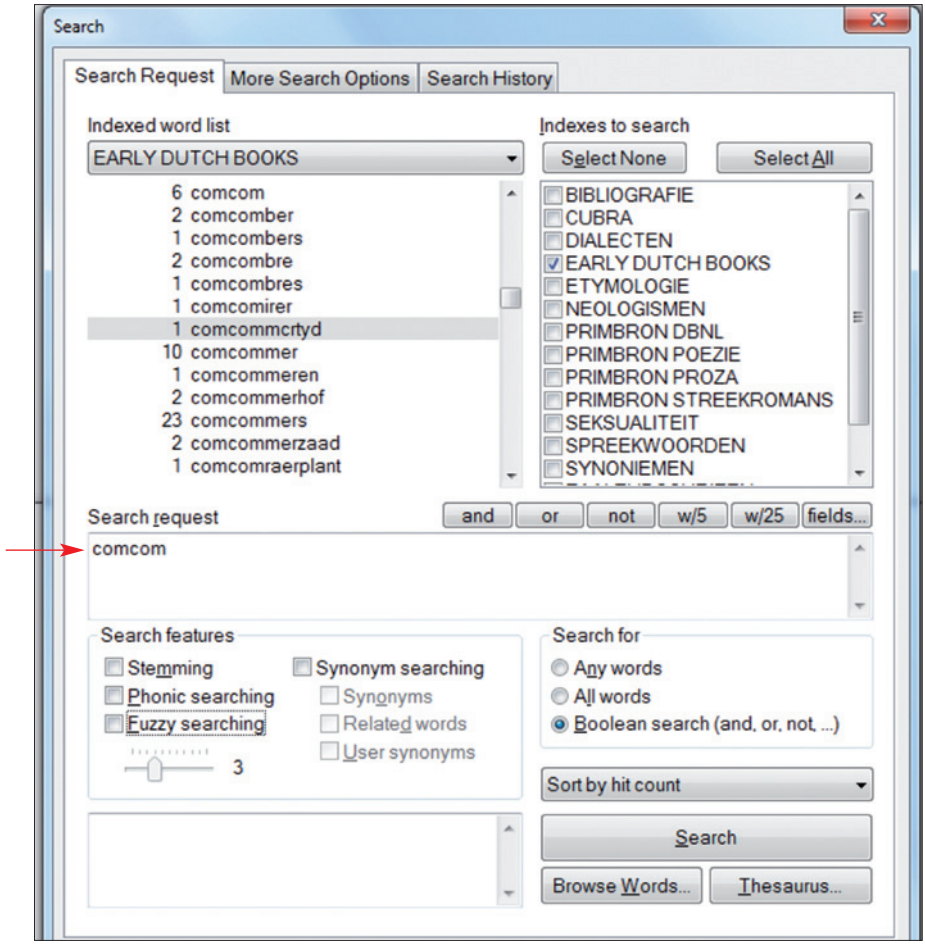
Stel, je wilt weten welke varianten er bestaan van de uitdrukking *met een aan zekerheid grenzende waarschijnlijkheid*. In de Historische Kranten vind je dat deze uitdrukking ontstond aan het begin van de 19de eeuw (oudste vindplaats 1829). Stond deze uitdrukking op zich of waren er indertijd vergelijkbare formuleringen in omloop?

Zoals gezegd zijn er op internet momenteel geen zoekmachines waarmee je woorden aan de voorzijde kunt trunceren, maar als je een website hebt laten indexeren met een webcrawler of je benadert een eigen verzameling bronnen met behulp van een indexeringstool, dan vind je het antwoord op deze vraag snel met bijvoorbeeld deze zoekopdrachten: **"*heid grenzende *heid"** en **"grenzende *heid"**

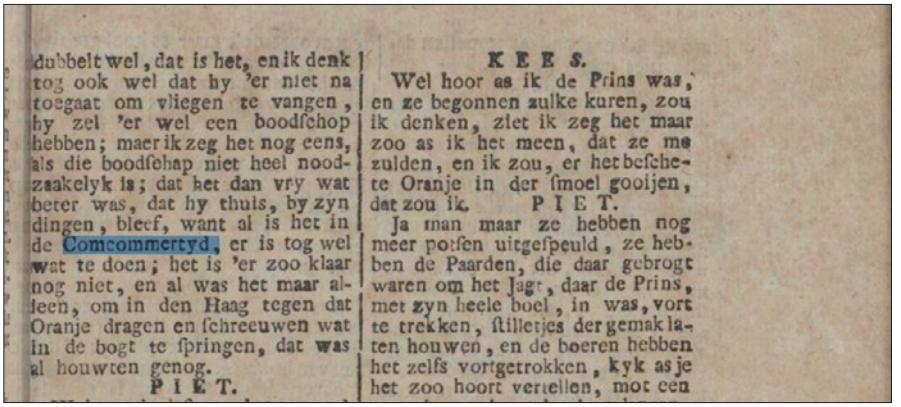
Resultaten:

- aan preutschheid grenzende fierheid (vroegste vindplaats 1816);
- aan waanzin grenzende droefheid (1832);
- aan wanhoop grenzende donkerheid (1836);
- aan onverschilligheid grenzende bedaardheid (1857);
- aan vermetelheid grenzende stoutmoedigheid (1868).

3. Met indexeringssoftware kun je zien welke ocr-fouten er in een bestand voorkomen. Nu moet je allerlei combinaties van wildcards verzinnen om dingen te vinden in onder meer Early Dutch Books Online (EDBO) en de Historische Kranten. Dat komt doordat je niet rechtstreeks in de index van die sites mag kijken. Als je via een webcrawler in bijvoorbeeld EDBO kijkt, of je zet die collectie op je eigen computer en je maakt er een index op, dan kan dit wel. Komt *komkommertijd* in EDBO wellicht nog in een andere vorm voor en zo ja, met welke ocr-fouten? Nou, zo bijvoorbeeld:



De vorm comcommcrtyd*, snel gevonden door in een eigen index op EDBO te kijken. Deze index is gemaakt met een tool genaamd dt-Search.*



dubbelt wel, dat is het, en ik denk tog ook wel dat hy 'er niet na toegaat om vliegen te vangen, hy zel 'er wel een boodſchop hebben; maer ik zeg het nog eens, als die boodſchap niet heel noodzaakelyk is; dat het dan vry wat beter was, dat hy thuis, by zyn dingen, bleef, want al is het in de Comcommertyd, er is tog wel wat te doen; het is 'er zoo klaar nog niet, en al was het maar alleen, om in den Haag tegen dat Oranje dragen en ſchreeuwen wat in de bogt te ſpringen, dat was al houwten genog.

PIET.

KEES.

Wel hoor as ik de Prins was, en ze begonnen zulke kuren, zou ik denken, ziet ik zeg het maar zoo as ik het meen, dat ze me zulden, en ik zou, er het beſchete Oranje in der ſmoel gooijen, dat zou ik.

PIET.

Ja man maar ze hebben nog meer potſen uitgeſpeuld, ze hebben de Paarden, die daar gebrogt waren om het Jagt, daar de Prins, met zyn heele boel, in was, vort te trekken, ſtilletjes der gemak laten houwen, en de boeren hebben het zelfs vortgetrokken, kyk as je het zoo hoort vertellen, mot een

Het Saturdags Kroegpraatje, 24 juni 1786.

Zoals in de inleiding al ter sprake kwam: voor sommigen zullen crawlers en indexeringstools te ver voeren, dus ik zal er hier niet over uitweiden. Maar met een relatief kleine investering kun je, in afwachting van al het prachtigs dat er nog aankomt, het slimme zoeken al naar The Next Level tillen.

Voor wie hiermee wil gaan experimenteren: de DBNL heeft een webpagina ter beschikking gesteld waar je diverse bronnenverzamelingen (datasets) kunt downloaden om mee aan de slag te gaan. Zie **www.dbnl.org/thema/zoek_ehbo.php**. Goede en betaalbare indexeringstools – met gratis testversies – zijn onder andere: dt-Search, Copernic search, X1 en (voor Apple-gebruikers) Foxtrot.

Er zijn twee tools die veel gemak kunnen opleveren: webcrawlers en indexeringssoftware. Webcrawlers zijn robots die websites indexeren. Indexeringstools doen hetzelfde met de gegevens op je computer. Voordeel: je kunt met deze tools nog slimmer zoeken dan nu op internet mogelijk is. Voor wie ermee wil experimenteren: bij de DBNL staan datasets klaar.

Slimmer documenteren

Tot slot: kun je met dergelijke indexeringstools alleen maar nóg slimmer zoeken? Nee, je kunt er ook slimmer mee documenteren.

Veel mensen gaan nu als volgt te werk. Van internet plukken ze allerlei bronnen voor hun onderzoek. Die zetten ze op hun pc, ze lezen ze of zoeken erdoorheen met Ctrl + f(ind), en ze maken aantekeningen. Als ze gaan schrijven, ordenen ze hun aantekeningen, gaan terug naar de bron, zoeken de passage, kijken of ze die nog steeds van belang vinden, enzovoorts.

Dit kan praktischer. In het voorbeeld hieronder gebruiken we een bron in pdf-formaat (zal het vaakst voorkomen), maar je kunt deze truc toepassen in alle bestandsformaten die door indexeringstools worden ondersteund. Let op: dit werkt natuurlijk alleen met bronnen die je op je eigen computer hebt gezet.

Je vindt in een pdf iets waarvan je denkt dat je het later kunt gebruiken. Je maakt je aantekening niet in een blocnote, Wordbestand of database, maar gewoon rechtstreeks op de pdf. Bijvoorbeeld zo:

Aan deze pdf is een notitie toegevoegd met Adobe Reader.

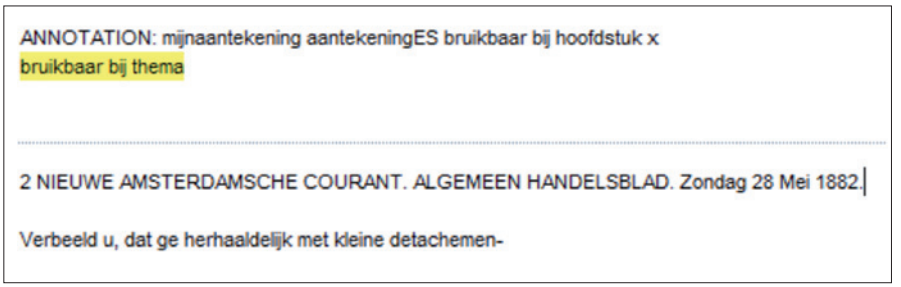
Of zo:

Op deze pdf is een passage gehighlight en is een notitie toegevoegd, beide met Adobe Reader.

Of je nu twee aantekeningen maakt of duizend, of je dit nu in één dag doet of gedurende vier jaar, het is handig je aantekeningen een redelijk consistente vorm te geven. Dat is zeker het geval als je die aantekeningen met meerdere mensen maakt. (Docent: 'Met deze werkgroep gaan we dertig romans lezen. Die gaan we analyseren op vijf thema's die jullie in vaste formuleringen op de pdf's moeten aanbrengen – zie het bijgevoegde lijstje').

Aan het eind van de rit is het niet nodig om pdf voor pdf te openen om je notities terug te lezen. In plaats daarvan zet je alle documenten bij elkaar in een map, die je vervolgens indexeert. Dit gaat supersnel (in sommige tools anderhalf miljoen woorden per minuut). Daarna kun je alle notities tegelijk doorzoeken (in feite heb je metadata toegevoegd, zoals dat in het jargon heet).

Snel alles bekijken over thema x, of alles bij elkaar vegen wat misschien in hoofdstuk y past? Je krijgt het in een net lijstje onder elkaar en je krijgt je notitie meteen te zien, zoals hier op de pdf van het *Algemeen Handelsblad* van 1882:



Een bijkomend voordeel van deze aanpak is dat je onderzoek herhaalbaar en controleerbaar is. 'Internet' is een beweeglijke bron: er komen voortdurend dingen bij, er vallen doorlopend dingen af. Of je een collectie bronnen nu 'verrijkt' met notities of alleen maar slim doorzoekt met een indexeringstool, of je alle bestanden nu in één map stopt of in een nette mappenstructuur, het doet er niet veel toe. Zodra je zo'n collectie – groot of klein – op je eigen computer hebt gezet, kun je zeggen: mijn conclusies zijn gebaseerd op deze bronnenverzameling. Wie ze wil controleren of wie mijn onderzoek wil overdoen, kan z'n gang gaan.

Wellicht zijn er lezers die denken dat dergelijke indexeringstools alleen geschikt zijn voor thuisgebruik, voor kleine verzamelingetjes, digitale peanuts.

Het is maar wat je peanuts noemt. Ik heb de afgelopen jaren vrij veel zelf gedigitaliseerd en ben actief gaan *harvesten*, zoals dat heet. Op mijn eigen pc staat een complete set van de Handelingen, van EDBO en van ruim 50.000 thematisch en chronologisch gerangschikte boeken en tijdschriften, waaronder veel van de DBNL. Alles bij elkaar – de websites die ik via een crawler benader niet meegeteld – gaat het om ruim negen miljoen pagina's, samen goed voor ruim vijf miljard woorden. Locatie: een externe harde schijf net iets groter dan een pakje sigaretten.

Het is waar, er komen nog veel grotere verzamelingen aan, maar in de tussentijd is dit een interessante verzameling om mee te onderzoeken of betaalbare en kant-en-klare indexeringstools goed werken voor verzamelingen van miljoenen pagina's. Mijn ervaring: als een tierelier. Ik werk al ruim drie jaar op deze manier en bijna dagelijks doe ik onverwachte vondsten op onverwachte plekken.

Soms door stom geluk, maar vaker door slim te zoeken.

Selectie uit de genoemde websites

Biografisch Portaal van Nederland, 20 biografische bronnen, ruim 70.000 levensschetsen
www.biografischportaal.nl

BNTL, Bibliografie van de Nederlandse Taal- en Literatuurwetenschap
www.bntl.nl

Ceneton, ruim 11.000 toneelteksten tot 1803
www.let.leidenuniv.nl/Dutch/Ceneton

DBNG, Digitale Bibliografie Nederlandse Geschiedenis
www.dbng.nl

Digitale Bibliotheek voor de Nederlandse Letteren, Nederlandstalige literatuur, taal en cultuurgeschiedenis van middeleeuwen tot nu
www.dbnl.org

Early Dutch Books Online, 11.240 boeken uit de periode 1781-1800
www.earlydutchbooksonline.nl

Etymologiebank, onmisbare bron voor geschiedenis Nederlandse woorden
www.etymologiebank.nl

Exalead, enige zoekmachine waarbij je woorden kunt afbreken (afbre*)
www.exalead.com/search

Geheugen van Nederland, ruim half miljoen items uit 102 collecties van 90 instellingen
www.geheugenvannederland.nl

Google Books, grootste bibliotheek uit de geschiedenis van de mensheid
www.books.google.nl of www.books.google.com

Google Scholar, grote verzameling wetenschappelijke publicaties
http://scholar.google.nl

Handelingen, Kamerverslagen, Kamerstukken en Kamervragen van 1814 tot 1 januari 1995
www.statengeneraaldigitaal.nl
Vanaf 1 januari 1995: https://zoek.officielebekendmakingen.nl/zoeken

Historische Kranten, uiteindelijk: 9 miljoen krantenpagina's van 1618 tot 1995
http://kranten.kb.nl

Meertens Instituut, databanken over o.a. namen, volkskundige publicaties en liederen
www.meertens.knaw.nl

Rollyo, maak snel je eigen zoekmachine
www.rollyo.com

Woordenboek der Nederlandsche Taal, beschrijft het Nederlands van grofweg 1500 tot 1976
http://gtb.inl.nl/?owner=WNT

Verder lezen

www.google.com/help/features.html en www.googleguide.com/advanced_operators.html

Zoekoperatoren voor Google

AND (of +)	= precies zo moet het er staan; dit moet voorkomen
OR	= geef me het ene OF het andere woord
NOT (of -)	= sluit dit woord uit
" "	[dubbele aanhalingstekens] = geef me exact deze woordgroep, in deze volgorde
*	[asterisk of sterretje] = vervangt ieder willekeurig woord
...	[drie puntjes] = geef me een waarde tussen x...y
~	[tilde] = zoek ook naar synoniemen
allintitle:	= alle gezochte woorden moeten in titel webpagina voorkomen
allinurl:	= alle gezochte woorden moeten in url voorkomen
define:	= geef de definitie van het gezochte woord
discussions:	= zoek in discussiegroepen, opiniepagina's etc.
filetype:	= zoek dit bestandsformaat. Bijvoorbeeld:
	doc of **docx** = MS Word
	pdf = Portable Document Format
	ppt of **pptx** = PowerPoint
	kmz = Google Earth
	xls of **xlsx** = Excel
inblogtitle:	= zoek in de titels van blogs
intitle:	= dit woord moet in de titel webpagina voorkomen
inurl:	= dit woord moet in url voorkomen
loc:	= zoek alles over deze locatie
movie:	= zoek informatie over films
related:	= zoek verwante, vaak buitenlandse sites
site:	= beperk de zoekopdracht tot deze website of dit domein. Bijvoorbeeld:
	edu = instellingen hoger onderwijs, o.a. universiteiten VS
	eu = Europese Unie
	gov = Amerikaanse federale overheid
	int = internationale organisaties (internationale verdragen)
	museum = musea
	org = non-profit organisaties

Enkele zoekoperatoren in Exalead

after:	= na een bepaalde datum (movie star after:2004/05/21)
before:	= voor een bepaalde datum (movie star before:2004/05/21)
soundslike:	= fonetisch
spellslike:	= fuzzy spelling

Voor een helder overzichtslijstje: www.exalead.com/search/web/search-syntax

Meer zoekoperatoren:

http://tinyurl.com/lijst1-operators

http://tinyurl.com/lijst2-operators

Geraadpleegde literatuur

- Henk van Ess, *De Google Code* (Amsterdam 2010)
- Wilfred de Feiter: *De beste gratis Google-toepassingen* (Culemborg 2004)
- Dirk de Grooff: *Goochelen met Google. Efficiënt leren zoeken met de grootste internetzoekmachine* (Leuven 2004)

Uitgave: Early Dutch Books Online, www.earlydutchbooksonline.nl
Tekst: Ewoud Sanders
Vormgeving: Tra[design], Nenke van Wermeskerken
Grafische productie: UFB/Grafimedia

Met dank aan George Boole, Isabel Brouwer, Josje Calff, Kurt De Belder, Esther ten Dolle, Jaap Engelsman, Henk van Ess, Chantal Keijsper, Nop Maas, Mathijs Sanders (geen familie), Jet Sanders (wel familie), Nicoline van der Sijs, Marijn Stockmann, Karin Swane, Rob Tempelaars en Rosalien van der Poel.

Dit boekje, een gezamenlijke uitgave van de universiteitsbibliotheken Leiden en Amsterdam en van de Koninklijke Bibliotheek, wordt u aangeboden door Early Dutch Books Online.

Oplage: 5.000

early dutch books ONLINE

Aanvullingen en correcties zijn welkom (via www.ewoudsanders.nl).
© Ewoud Sanders, 2011

De publicatie Eerste Hulp Bij e-Onderzoek is ook in open access beschikbaar: http://hdl.handle.net/1887/17774