

Digital Humanities

Een digitale lens op de producten van de menselijke geest

MARIJN KOOLEN

1 Definitie

Digitale geesteswetenschappen, beter bekend als Digital Humanities, laten zich moeilijk definiëren. De afgelopen jaren hebben veel onderzoekers geprobeerd een goede definitie te geven¹, maar de kern is vaak hetzelfde en die draait om het gebruiken van computers en software voor het (semi)automatisch analyseren en presenteren van geesteswetenschappelijk onderzoek.

Traditioneel maken de meeste geesteswetenschappelijke disciplines voornamelijk gebruik van kwalitatieve onderzoeksmethoden. Kritische reflectie en interpretatie zijn vaak de methoden die geassocieerd worden met traditioneel geesteswetenschappelijk onderzoek. Producten van de menselijke geest worden zoveel mogelijk in originele vorm bestudeerd. Bij digitale geesteswetenschappen ligt de nadruk meer op kwantitatieve methoden: tekstueel, auditief en visueel materiaal wordt opgebroken in kleine eenheden die geteld kunnen worden. Door digitalisering kan materiaal gekwantificeerd worden, waarmee andere analyses mogelijk zijn dan wanneer het materiaal in zijn geheel en in een bredere context wordt beschouwd.

Volgens sommigen is het een meervoudsterm, de digitale geesteswetenschappen, omdat het om een meervoud van nieuwe methoden en perspectieven gaat, zonder dat er een duidelijk eigen vakgebied is. Ze bestrijken alle disciplines van de geesteswetenschappen, maar hebben geen eigen onderwerp van studie. Wat onderzoekers in Digital Humanities bij elkaar brengt is de vergelijkbare technische problematiek tussen

verschillende geesteswetenschappelijke disciplines en de gedeelde waarden over de zin en onzin van digitale onderzoeksmethoden in een vakgebied dat traditioneel vooral leunt op interpretatie en kritische reflectie.

Een ander, veelvoorkomend perspectief is dat digitale geesteswetenschappen slechts een tijdelijk label is, en het woord *digitale* zal wegvallen zodra digitale onderzoeksmethoden niet meer als nieuw worden ervaren, en het een geaccepteerd onderdeel is geworden van alle geesteswetenschappen.

Een van de belangrijke redenen dat Digital Humanities (DH) als een verzamelnaam wordt gebruikt voor zeer uiteenlopend digitaal onderzoek in de geesteswetenschappen is de snel groeiende hoeveelheid digitale data en het gebrek aan goede middelen om hier mee om te gaan. Deze situatie symboliseert een soort *informational turn* die eerder in andere vakgebieden heeft plaatsgevonden, zoals natuurkunde, biologie, scheikunde, astronomie, economie en de sociale wetenschappen. De data wordt niet gemaakt of verzameld naar aanleiding van een onderzoeksvraag, maar juist in tegengestelde richting worden onderzoeksvragen bedacht rondom de enorme berg data die beschikbaar is (Gray, 2007). DH-onderzoek is vaak data-gestuurd en exploratief. Men zoekt naar interessante patronen, structuren en relaties die dan door grondige analyse en met kritische reflectie geïnterpreteerd moeten worden.

Ondanks of juist dankzij deze diverse definities van Digital Humanities krijgt dit onderwerp steeds meer aandacht en wagen steeds meer geesteswetenschappers en onderzoekers uit andere disciplines zich aan de vele en vaak complexe digitale onderzoeksmethoden. Maar de explosieve groei in de afgelopen jaren wordt vooraf gegaan door een lange geschiedenis.

2 Geschiedenis

Digital Humanities of de gekwantificeerde benadering in de geesteswetenschappen is niet nieuw, maar kent een lange ontwikkeling, en het huidige onderzoek is onderdeel van een lang traject waarin methoden zijn bedacht om inzichten te krijgen in geesteswetenschappelijke kwesties door middel van digitalisering.²

Al rond 1260 werd een eerste concordantie van de Bijbel gemaakt, waarin alle woorden in een alfabetisch geordende index werden opgenomen, met per woord de context. i.e. de direct voorafgaande en volgende woorden uit de tekst. Door de eeuwen heen zijn van verschillende versies van de Bijbel concordanties gemaakt, en elk project kostte vele jaren mensenwerk.

Joseph Scaliger gebruikte rond 1600 kwantitatieve methoden om kalenders van verschillende historische perioden en gebieden te vergelijken en bewees daarmee dat de oude Egyptenaren leefden voor de gebeurtenissen in het Oude Testament.

Ook Stylometrie, de kwantitatieve analyse van schrijfstijl, ontstond ver voor de uitvinding van de elektronische computer. Een van de bekendere voorbeelden is gerelateerd aan de discussie rond het auteurschap van de werken van Shakespeare, die speelde halverwege de 19de eeuw. Aan het einde van die eeuw vergeleek T.C. Mendenhall de stijl van verschillende auteurs, o.a. Shakespeare, Francis Bacon en Christopher Marlowe, door te tellen hoe vaak elk van deze auteurs specifieke woorden van verschillende lengte gebruikte.

Kort na de uitvinding van de elektronische computer kwamen wetenschappers in allerlei gebieden op het idee om de rekenkracht van zulke computers in te zetten voor onderzoek. Geesteswetenschappers waren hierop geen uitzondering.

Het Begin: Humanities Computing

Het meest genoemde beginpunt van digitale geesteswetenschappen is het werk van Roberto Busa. In 1949 startte hij een project om een index te maken van het volledige werk, 11 miljoen woorden in totaal, van Thomas van Aquino. Hij benaderde IBM om het proces deels te automatiseren. Alle teksten werden opgeslagen op ponskaarten, en er werd een concordantieprogramma gemaakt waarmee men elk woord op lemma kon opzoeken, en zien hoe vaak en waar elk woord voorkomt. Ondanks het feit dat IBM hun medewerking verleende duurde het tot 1974 voor de eerste gedrukte editie verscheen (Hockey, 2004).

De elektrische computer werd in eerste instantie voornamelijk gebruikt voor concordanties en auteursattributie (het identificeren van de auteur

van een stuk tekst). Dit soort onderzoek werd in eerste instantie *Humanities Computing* genoemd. De grootste problemen waar men mee geconfronteerd werd, waren het identificeren van woorden en het omgaan met linguïstische variatie, zoals woordvervoegingen, synoniemen en homoniemen. Dit zijn nog steeds belangrijke problemen voor bovengenoemde onderzoeksvormen, al zijn er inmiddels computerprogramma's om een deel van de variatie te normaliseren. De meeste programma's zijn ontwikkeld voor het Engels, maar voor veel Europese talen zijn er ook (deel)oplossingen.

Van Mainframe Naar PC

Voor de introductie van de Personal Computer (PC) maakten wetenschappers gebruik van zogenaamde *mainframes*. Deze centrale computers werden gedeeld door vele wetenschappers, die rekentijd konden aanvragen. Vaste ondersteuners kregen vele verzoeken van vergelijkbare aard, en om rekentijd en opslag efficiënt te gebruiken werden generieke programma's ontwikkeld die door vele iteraties steeds preciezer en effectiever werden.

Een van de gevolgen van het toenemende gebruik van pc's was dat wetenschappers vaker zelf software ontwikkelden en digitale experimenten deden, waardoor minder gebruik werd gemaakt van de opgebouwde expertise van mainframe ondersteuners, die daardoor ook afnam. En er werd meer dubbel werk gedaan op verschillende plekken. Daarentegen kon men meer interactief met software experimenteren en werden er meer innovatieve oplossingen gevonden (Hockey, 2004). Veel DH pioniers verdiepten zich meer en meer in de technische mogelijkheden en beperkingen en werden zelfstandiger.

Het voorkomen van duplicatie van werk wordt in onderzoeksprojecten deels opgelost door samen te werken in interdisciplinaire teams van GW onderzoekers, informatiespecialisten en programmeurs, en gebruik te maken van ieders kennis uit andere projecten (zie Sectie 4). Een andere belangrijke manier om duplicatie te voorkomen is het delen van data en kennis via het Web en het zoveel mogelijk met open source producten werken.

Text Encoding

Aan de ontwikkeling van *markup talen* -waarmee documenten gestructu-

reerd worden door beschrijvende metadata, zoals in HTML en XML- hebben geesteswetenschappers een belangrijke bijdrage geleverd. De toename van het aantal beschikbare digitale teksten creëerde een behoefte aan standaard bestandsformaten en coderingsschema's zodat men materialen kon uitwisselen. Tot aan 1986 maakten verschillende tekstanalyseprogramma's gebruik van verschillende, conflicterende coderingsschema's, waardoor digitale teksten vaak van het ene in het andere formaat overgezet moesten worden (Hockey, 2004). Met de publicatie van de *Standard Generalized Markup Language* (SGML) door de ISO vonden taal- en literatuurwetenschappers een *markuptaal* waarmee verschillende soorten tekst en metadata in een gemeenschappelijk coderingsschema uitgedrukt konden worden.

```
<sp>
<speaker>Prospero</speaker>
<l part="Y">I'll deliver all,</l>
<l>And promise you calm seas, auspicious gales,</l>
<l>Be free and fare thou well. <stage type="exit">Exit Ariel</stage>
  Please you, draw near. <stage type="exit">Exeunt all but Prospero</stage>
  <note place="margin">Epilogue</note>
</l>
<l>Now my charms are all o'erthrown,</l>
<l>And what strength I have's mine own</l>
<l>As you from crimes would pardoned be,</l>
<l>Let your indulgence set me free.</l>
</sp>
<stage type="mix">He awaits applause, then exit.</stage>
```

bron: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DR.html>

Hieruit is het *Text Encoding Initiative* (TEI, figuur 1) ontstaan. Deze richtlijnen hebben bijgedragen aan de ontwikkeling van XML en zijn inmiddels ook gedefinieerd in XML formaat. Toen de behoefte aan beschrijvende metadata op het Web toenam, zocht men naar een nieuwe, flexibele *markuptaal* voor de vele soorten tekstdocumenten en beschrijvingen. Het door geesteswetenschappers ontwikkelde TEI bleek een grote inspiratiebron voor wat nu een van de meestgebruikte dataformaten voor het uitwisselen van gegevens is.

Beschrijvende *markup* heeft een aantal grote voordelen voor digitaal GW-onderzoek. Ten eerste worden opslag en retrieval makkelijker doordat documenten als database bevraagd kunnen worden en de *markup* gebruikt kan worden om door lange documenten te navigeren. Bovendien kunnen documenten door middel van *markup* automatisch geanalyseerd worden. Onderdelen van de tekst als paragrafen, zinnen,

regels, uitdrukkingen, woorden en verwijzingen kunnen apart gemarkeerd en onderscheiden worden (Renear, 2004). Met de komst van het Web werden veel collecties met gedigitaliseerde teksten online beschikbaar gemaakt in tekstarchieven.

Van Humanities Computing naar Digital Humanities

Sinds de komst van het World Wide Web zijn digitale collecties en analyseprogramma's beter toegankelijk geworden. Hierdoor zijn langzaam maar zeker meer onderzoekers in aanraking gekomen met de mogelijkheden van digitale onderzoeksmethoden. Er ontstond ook een groep die zelf geen digitale technieken gebruikte maar kritisch reflecteerde op die nieuwe middelen. Er ontstond een splitsing tussen GW onderzoekers die technieken en datasets creëren en onderzoekers die DH meer theoretisch benaderen. De toenemende zichtbaarheid van DH leidde tot het opnemen van DH in academische programma's. De scope werd breder, digitale presentatie werd ook gezien als DH, en computatie stond steeds minder centraal. Projecten werden meer interdisciplinair en de gemeenschap werd steeds heterogener. Men beseftte dat er geen centrale methode of onderwerp was, maar dat DH als gemeenschap vooral bepaalde waarden en perspectieven deelde. Langzaam maar zeker werd de term *humanities computing* vervangen door *digital humanities*. Sommigen hoopten hiermee de gemeenschap inclusiever en opener te maken. Of de naamsverandering gewerkt heeft is moeilijk vast te stellen, maar het is duidelijk dat met deze verandering de bekendheid en reputatie van Digital Humanities enorm zijn toegenomen.

3 State Of The Art

Deze sectie geeft eerst een overzicht van onderzoeksvragen en -projecten binnen de digitale geesteswetenschappen en behandelt daarna de meestgebruikte digitale methoden en tools om die vragen te kunnen beantwoorden. Veel van de voorbeelden zijn gebaseerd op tekstuele bronnen, omdat daar het meeste digitale gereedschap voor is. Er zullen echter ook projecten en tools voor niet-tekstueel materiaal aan bod komen. Uiteraard is dit geen volledig overzicht van alle relevante

projecten. Er is gepoogd om een diverse groep, voornamelijk Nederlandse projecten te bespreken uit verschillende GW disciplines, waarin verschillende technieken gebruikt worden. De technieken die in de projectenbeschrijvingen genoemd worden, worden verderop toegelicht.

Huidig DH Onderzoek

Archeologie

Archeologen gebruiken digitale methoden om historische gebouwen digitaal te reconstrueren, zoals in het 3/4D Research Lab.³ Aan de hand van tekstuele beschrijvingen en schetsen van een gebouw kan een reconstructie worden gemaakt. Een van de vele interessante aspecten van digitale reconstructie, waarbij rekening wordt gehouden met de eigenschappen van bouwmaterialen, is dat nauwkeurige modellering soms aan het licht brengt dat bepaalde historische beschrijvingen en schetsen niet kunnen kloppen, waarmee nieuwe inzichten verkregen worden en huidige kennis over historische bouwwerken en stijlen herzien kan worden.

Deze reconstructies kunnen worden gebruikt in virtuele 3D werelden, in combinatie met narratieven, om historische locaties en gebeurtenissen tot leven te brengen, in bijvoorbeeld virtuele musea, zoals in het project V-MusT.⁴

Film- en Televisiewetenschappen

In het project *Emotions in Film*⁵ wordt onderzocht hoe mensen emotioneel reageren op bepaalde filmscènes. De emotionele reacties worden verkregen zowel in de vorm van tekst, uit online recensies en discussies over beroemde scènes op blogs en forums, als in gezichtsuitdrukkingen. De gezichtsuitdrukkingen van kijkers worden geregistreerd tijdens testsessies waarin dezelfde scènes vertoond worden en vervolgens geanalyseerd op zeven emotiecategorieën: blij, boos, verrast, bang, bedroefd, verafschuwd en neutraal. Dit wordt automatisch gedaan door middel van een zogenaamde *facereader* tool.⁶ Hiervoor is een *machine learning* algoritme gebruikt om emoties te leren herkennen in gezichtsuitdrukkingen.



Filosofie

Het Europese project Tarski's Revolution⁷ ontwikkelt digitale tools en methodologieën om de ideeëngeschiedenis in kaart te brengen en te traceren hoe concepten zich verspreiden en veranderen. Technieken als *text encoding (TEI)*, *computationele linguïstische analyse* en *Linked Open Data* worden gebruikt om filosofische teksten te annoteren, de verschillende concepten te relateren (linken) aan elkaar en aan personen, perioden en geografische locaties. Hiermee wordt een kwantitatieve methodologie ontwikkeld voor het analyseren van de geschiedenis van concepten.

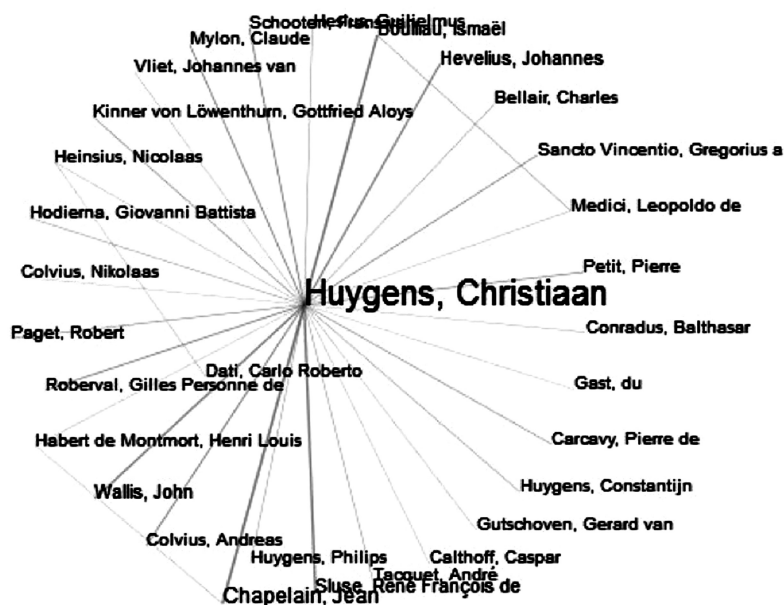
Geschiedenis

Het project *Circulation of Knowledge and Learned Practices in the 17th-century Dutch republic*⁸ richt zich op de netwerken van wetenschappers in de 17de eeuw en hoe nieuwe kennis zich via correspondenties heeft verspreid. Een collectie van 20.000 brieven van Nederlandse 17de-eeuwse wetenschappers en hun internationale collega's is geannoteerd met namen van de zender, ontvanger en personen die genoemd worden, hun wetenschappelijke rol, evenals de belangrijkste onderwerpen van de brieven. Deze gestructureerde informatie wordt digitaal als netwerk geanalyseerd en gevisualiseerd om te achterhalen wie bijvoorbeeld de centrale personen waren in het ontdekken van de ringstructuur en manen rond Saturnus.⁹

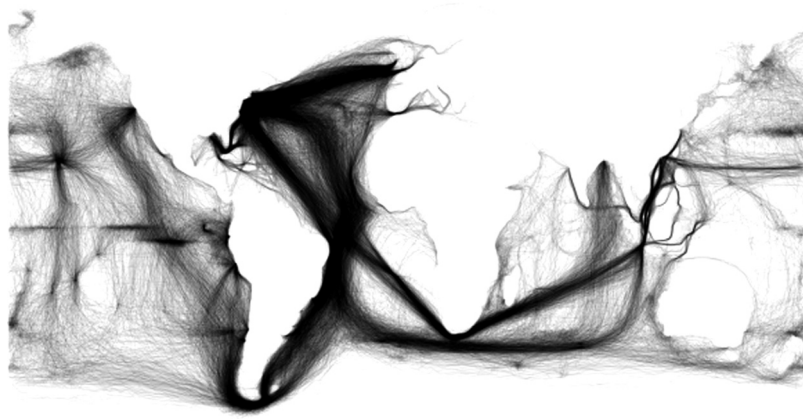
IV. PERIODE 1654-1659

ePistolarium: zoekwoord: 'Saturn' – 155 brieven

Correspondenten:



Ben Schmidt heeft een grote collectie van gedigitaliseerde scheepslogboeken gebruikt om te visualiseren hoe de scheepvaart van Europa en Amerika zich ontwikkeld heeft, hoe verschillende routes gebruikt werden en hoe met name de walvisvaart veranderd is.¹⁰ De logboeken bevatten miljoenen datapunten in de vorm van geografische coördinaten waarmee zeereizen gevisualiseerd kunnen worden.



Het project *Time Capsule*¹¹ linkt historische farmaceutische en botanische documenten en kennis uit heterogene bronnen, zodat ze op eenduidige manier bevestigd, geanalyseerd en gevisualiseerd kunnen worden. De gegevens uit de verschillende soorten bronnen worden geannoteerd met temporele en geografische gebonden perspectieven, waarmee zogenaamde tijdschapsules (time capsules) ontstaan waarin automatisch, door middel van *Linked Open Data technieken* verbanden kunnen worden ontdekt.

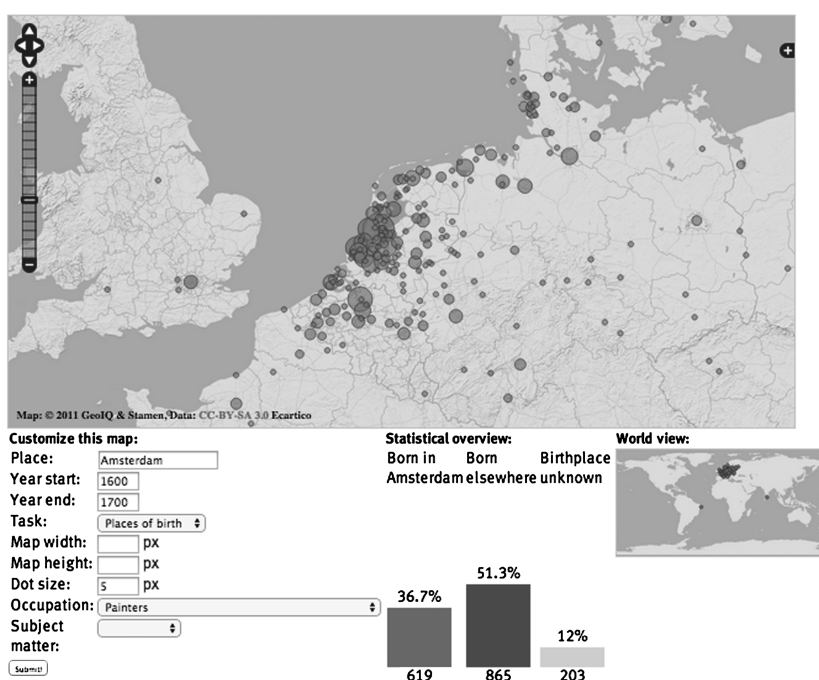
*Translantis*¹² onderzoekt de intrede van de Verenigde Staten van Amerika in de publieke discours in Nederland. Hiervoor worden teksten op woordniveau geannoteerd en geanalyseerd. Door middel van automatische naamherkenning (*Named Entity Recognition*) en *topic modelling* kan inzichtelijk worden gemaakt wanneer, in welke media en op wat voor manier, de VS besproken werden.

Kunstgeschiedenis

*De culturele industrie van Amsterdam in de Gouden Eeuw*¹³ is een project dat onderzoekt hoe Amsterdam als een cultureel centrum in Europa zich ontwikkeld heeft, door de complexe relaties tussen kunstenaars, schrijvers, uitgevers, handelaars en verzamelaars in kaart te brengen. Gegevens over deze personen, zoals geboorte-, trouw- en overlijdensgegevens, woon- en werkadressen en persoonlijke en professionele

relaties worden verzameld in de ECARTICO database,¹⁴ waarmee vervolgens geografisch-temporele visualisaties kunnen worden gemaakt, en de sociale netwerken geanalyseerd kunnen worden (zie *netwerkanalyse* in de volgende sectie).

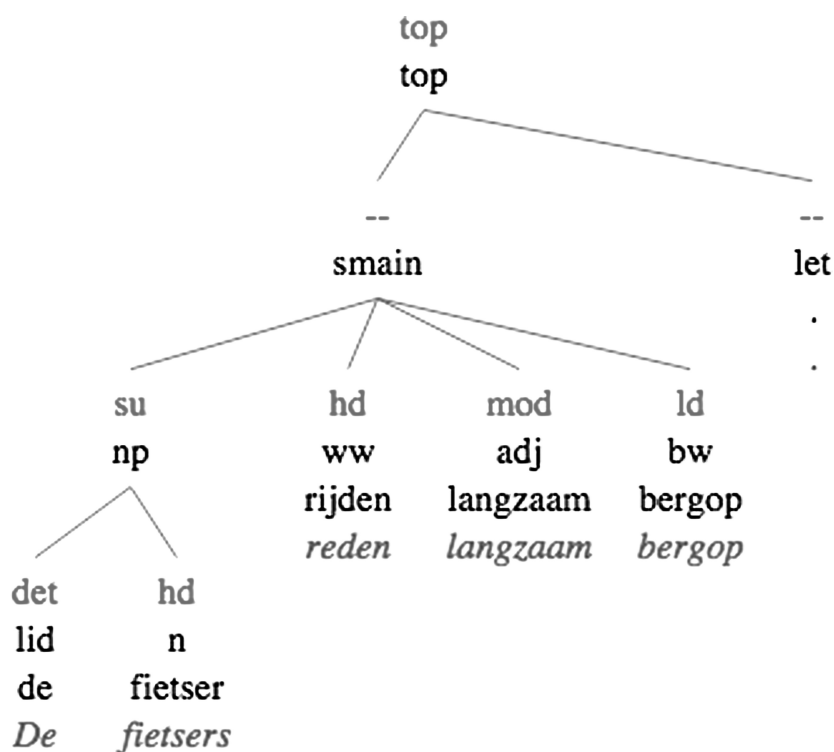
Places of birth of painters active in Amsterdam between 1600 and 1700



Literatuurwetenschappen

Wat maakt een boek literair? Deze vraag staat centraal in het project *The Riddle of Literary Quality*.¹⁵ De onderzoekers hebben via een groot nationaal lezersonderzoek gegevens verzameld over welke boeken mensen gelezen hebben, of ze die goed of slecht vinden en hoe 'literair' ze die vinden.¹⁶ Een van de vroege resultaten van dit project komt uit een onderzoek van Jautze et al. (2013) waarin ze wilden weten in welke opzichten *Chicklit* (humoristische eigentijdse romans met als doelgroep jonge vrouwen) en zogenaamde 'serieuze' literatuur van elkaar verschillen. Ze gebruikten software voor het identificeren van zinnen en het

ontleden van de zinsstructuur. Literatuur heeft over het algemeen langere en complexere zinnen, e.g. meer bijzinnen.



Tangherlini & Leonard (2013) onderzochten de invloed van Darwin's *On the Origin of Species* op Deense literatuur met behulp van *topic modeling* (uitgelegd in de volgende sectie). Hiervoor hebben ze software gebruikt die automatisch bepaalt in hoeverre een document (roman, nieuwsbericht, brief of politieke discussie) over een specifiek topic gaat.

Kestemont et al. (2013) hebben een prachtige wetenschappelijke documentaire¹⁷ gemaakt over hun stylometrie-onderzoek naar het werk van de 12de-eeuwse Hildegard van Bingen, waarin de vraag centraal staat hoe kopiïsten de schrijfstijl van Hildegard van Bingen hebben beïnvloed.

Musicologie

In musicologie zijn ook veel projecten, waarin onderzoeksvragen aan de orde komen als: Wanneer lijken twee stukken muziek op elkaar? Hoe

verspreidden muziekstukken zich in Europa in de 16de en 17de eeuw? Wat maakt een liedje ‘catchy’? En waarom blijven sommige melodieën eendeloos in je hoofd hangen (‘oorwurmen’) terwijl andere meteen vergeten zijn?

Honingh en Bod (2012) hebben onderzocht hoe de frequentie van toonsafstanden (pitch intervals) tussen twee opeenvolgende tonen gebruikt kunnen worden om de stukken van Beethoven te classificeren in drie verschillende perioden. Dezelfde methode werkt ook om genres als jazz, rock en pop van elkaar te onderscheiden. In een ander project heeft Honingh d.m.v. experimenten met luisteraars onderzocht wanneer twee muziekstukken op elkaar lijken.¹⁸ In sommige stukken is het ritme leidend en bepalend of een ander stuk erop lijkt. In andere stukken kan de melodie een grotere rol spelen. Deze kennis wordt gebruikt om applicaties te ontwikkelen die dj’s kunnen helpen bij het vinden van muziekstukken die goed in elkaar over kunnen lopen.

Het COGITCH¹⁹ project onderzoekt cognitieve aspecten van muziek, zoals hoe oorwurmen ontstaan en welke aspecten een liedje *catchy* maken. Hierin worden muziekfragmenten niet alleen automatisch geanalyseerd in termen van bijvoorbeeld tempo, ritme, melodie, etc., maar wordt vooral ook onderzocht hoe luisteraars muziek ervaren. Met behulp van een online spel, Hooked!, worden gegevens verzameld over welke fragmenten van liedjes mensen snel kunnen herkennen, onthouden en in gedachten kunnen reproduceren, waarna fragmenten ontleed worden om inzicht te krijgen in wat sommige fragmenten *catchy* maakt en andere niet.

In het Tunes & Tales²⁰ project onderzoekt men orale transmissie van volksverhalen en volksliedjes. Door verschillende overgeleverde varianten met elkaar te vergelijken kunnen eerdere versies afgeleid worden en modellen ontwikkeld worden die kunnen verklaren hoe en waarom verhalen en liederen veranderen, geografisch en door de tijd. Daarnaast kan met een grote verzameling van liederen en verhalen en al hun varianten inzicht verkregen worden waarom sommige elementen van een verhaal of lied stabiel blijven terwijl andere veranderen. In het project worden verhalen en liederen als gelaagde sequenties van motieven gemodelleerd. Op die manier kunnen algoritmes ontwikkeld worden om automatisch motieven te herkennen en te identificeren.

Bovendien kunnen algoritmes voor het meten van similariteit gebruikt worden voor muziekzoekmachines, zoals in het onderzoeksproject WITCHCRAFT²¹, waarin de gebruiker een muziekstuk als voorbeeld invoert, of een melodie neuriet en de software een lijst resultaten geeft van vergelijkbare muziek.

Nieuwe Media

Hoe is de structuur van de Nederlandse *blogosphere* (de gemeenschap van bloggers die naar elkaars berichten verwijzen en daarop reageren) door de jaren heen veranderd (Weltevrede en Helmond, 2012)? Het Internet Archive verzamelt sinds 1996 historische versies van zoveel mogelijk webpagina's. Weltevrede en Helmond hebben met lijsten van URL's van Nederlands blogs gezocht in historische pagina's van Nederlandse bloggers en hebben de hyperlinkstructuur tussen die blogs geanalyseerd. Door het netwerk te visualiseren en te analyseren konden zij zien dat Nederlandse bloggers verschoven van eigen sites in het .com domein naar standaard blogplatforms en sites binnen het .nl domein.

Wat is de impact van sociale media op de openbare orde? Hoe hebben sociale media bijgedragen aan de problemen rondom het zogenaamde Project X en het uit de hand gelopen verjaardagsfeestje in Haren?²² Om inzicht hierin te krijgen, zijn tienduizenden relevante berichten van Facebook en Twitter verzameld en geanalyseerd. Gegevens als leeftijd, geslacht en locatie van de gebruikers en tijdstip van de berichten laten zien welke groepen gebruikers betrokken waren en hoe zij elkaar beïnvloeden.

Taalkunde

In de taalkunde zijn digitale onderzoeksmethoden al decennialang gevestigd. Er zijn vele onderzoeksprojecten geweest.

Een voorbeeld van een digitaal taalkundig onderzoeksproject is *Gabmap*,²³ waarin variaties in dialecten geanalyseerd en gevisualiseerd worden. Tabellen met gegevens over de uitspraak van woorden en de geografische gegevens van de sprekers kan worden ingevoerd in de Gabmap applicatie, waarmee de data op verschillende manieren gegroepeerd (*clustering*) kan worden en bijvoorbeeld regionale verschuivingen en overeenkomsten in dialecten zichtbaar gemaakt kunnen worden.

*Monk*²⁴ is een langlopend project om handgeschreven bronnen in archieven automatisch doorzoekbaar te maken. OCR technieken zijn met name ontwikkeld voor gedrukte tekst, waarbij veelal vaste lettertypen worden gebruikt in een document. De variatie in handgeschreven letters is veel groter, waardoor correcte letterherkenning een grotere uitdaging is. Door middel van handmatig geannoteerde voorbeelden en *machine learning* technieken leert het Monk systeem gaandeweg steeds beter en steeds meer handgeschreven tekst te herkennen. Hiermee worden deze bronnen geschikt voor *information retrieval* technieken die in zoekmachines worden gebruikt.

Ondersteunende Projecten

Er is ook een aantal ondersteunende projecten, met name gericht op het opzetten van een gedeelde infrastructuur waarin data, software en resultaten gedeeld kunnen worden en waarin onderzoekers kunnen samenwerken.

CLARIAH (Common Lab Research Infrastructure for the Arts and Humanities) is een Nederlands consortium van meer dan 40 partners, met als doel het bevorderen van een infrastructuur voor DH projecten, zodat digitale datasets en tools makkelijk ontwikkeld en toegankelijk gemaakt kunnen worden. Hiermee beoogt men om de drempel te verlagen om DH onderzoek te doen, en geesteswetenschappers een beter inzicht te bieden in wat er mogelijk is. Tegelijkertijd kan hiermee in kaart worden gebracht wat de wensen zijn voor zo'n digitale onderzoeksinfrastructuur. Op Europees niveau is hiervoor DARIAH (Digital Research Infrastructure for the Arts and Humanities),²⁵ wat eenzelfde doel voor ogen heeft. De Nederlandse tak hierbinnen, DARIAH-NL, fungeert als een brug tussen DARIAH-EU en CLARIAH.

Het NWO programma *CATCH*²⁶ (*Continuous Access To Cultural Heritage*) is een langlopend project waarin informatica-onderzoekers samenwerken met erfgoedinstellingen om gedigitaliseerd erfgoedmateriaal toegankelijk te maken voor zowel het algemene publiek als voor onderzoek. In een groot aantal projecten zijn digitale collecties en tools ontwikkeld die direct relevant zijn voor digital humanities onderzoek. Het eerder genoemde project WITCHCRAFT is een van de projecten binnen CATCH.

Methoden, Technieken en Tools

Er zijn veel verschillende computationele technieken die gebruikt worden in Digitaal GW-onderzoek. Hier volgt een kort overzicht van de meestgebruikte, meest generieke technieken en methoden in Digital Humanities onderzoek. Bij elke techniek wordt een aantal onderzoeksprojecten en softwareprogramma's als voorbeeld genoemd.

Topic modeling algoritmes kunnen detecteren welke groepen woorden relatief vaak samen voorkomen, waarbij elke groep woorden een specifiek onderwerp (topic) vertegenwoordigt. Dit wordt automatisch afgeleid uit een groot corpus van teksten. Met deze topics kan vervolgens gemeten worden in hoeverre een document een topic behandelt. Dit is met name interessant om een voorkomen van topics door de tijd heen te analyseren (bijv. in chronologisch geordende collectie zoals een dagboek of een archief) of twee of meer documentencollecties te vergelijken. Deze techniek wordt in veel verschillende projecten gebruikt. Meeks en Weingart (2012) geven een goed overzicht van de rol en het belang van *topic modeling* in Digital Humanities onderzoek.

- *Mallet*²⁷ is een van de meestgebruikte topic modelingprogramma's binnen de digitale geesteswetenschappen. Het programma bevat ook modules voor het clusteren (automatisch groeperen) van documenten en het extraheren van linguïstische sequenties (zoals herkenning van namen in teksten).
- *GenSim*²⁸ is een topic modeling module voor Python programmeurs. Zij kunnen deze module inladen in hun programma's en daarmee topic modellen maken voor allerlei soorten gegevens. Het grote voordeel van deze programmatische aanpak is dat de gebruiker complete controle heeft over de eenheden waarmee topic modellen gemaakt worden (bijv. woorden maar ook sequenties van woorden of geografische coördinaten).

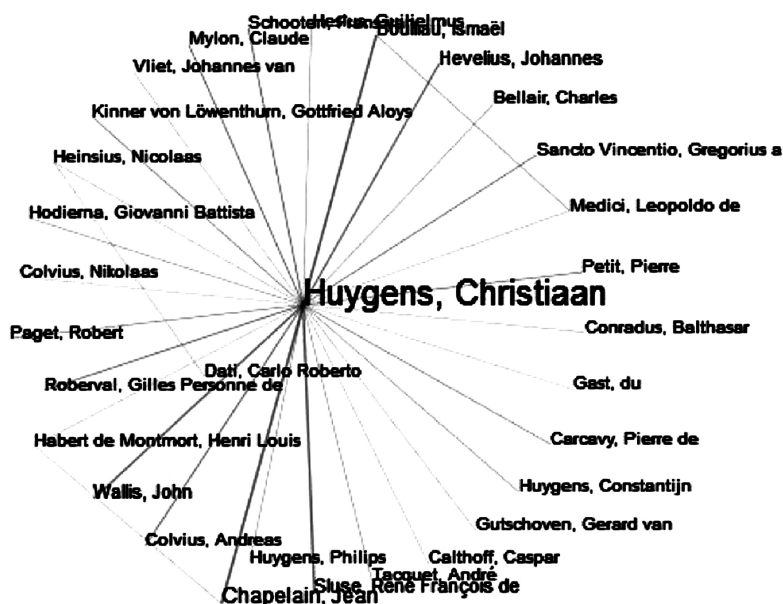
Netwerkanalyse is interessant voor het analyseren van gestructureerde (meta) data. Uit dit soort data kunnen relaties worden afgeleid tussen entiteiten zoals personen, objecten, locaties, etc. De structuur kan wiskundig geanalyseerd worden om bijvoorbeeld te bepalen welke entiteit het meest centraal is of welke entiteiten dicht verbonden clusters vormen. Deze entiteiten en relaties kunnen als een netwerk worden weergegeven, waarmee bijvoorbeeld de belangrijkheid of rol van een persoon zichtbaar wordt. Een voorbeeld hiervan is het netwerk van wetenschappers in Europa in de 17^e

eeuw die met elkaar correspondeerden over de ringstructuur van manen rond Saturnus. De briefwisselingen kunnen grafisch worden weergegeven, zowel op een geografische kaart, als in een netwerk van personen. Door het netwerk van correspondenten te visualiseren op basis van brieven in verschillende periodes, wordt zichtbaar wie in elke periode de belangrijkste wetenschappers waren, en hoe het debat zich ontwikkelde.²⁹

IV. PERIODE 1654-1659

ePistolarium: zoekwoord: 'Saturn' – 155 brieven

Correspondenten:



Een ander voorbeeld is een netwerkanalyse van de interrupties in de Tweede Kamer. De politieke handelingen in de Tweede Kamer worden nauwkeurig vastgelegd in gestructureerde data. Zo kan automatisch achterhaald worden hoe vaak sprekers onderbroken worden door leden van andere partijen, en bij welke onderwerpen.

Een laatste voorbeeld is het analyseren van de netwerkstructuur van webpagina's en sociale media.

- *NodeGoat*³⁰ is een programma voor data management, analyse en visualisatie waarmee geografische en abstracte netwerken van entiteiten

en relaties gemaakt kunnen worden. Als de ingevoerde gegevens ook temporele data bevat (bijv. jaartallen), kunnen de verschuivingen in het netwerk door de tijd heen ook gevisualiseerd worden.

- *Gephi*,³¹ *VOSviewer*³² en *Payek*³³ zijn generieke programma's voor netwerkanalyse en -visualisatie. De analyse betreft de wiskundige eigenschappen van netwerken die inzicht geven in de verhoudingen tussen entiteiten als de relatieve prestige en status van entiteiten. Entiteiten die verbonden zijn met veel andere entiteiten zijn heel zichtbaar en centraal, en daarmee belangrijker voor het netwerk dan entiteiten die nauwelijks of niet verbonden zijn met andere.

Linguïstische analyse tools worden gebruikt om tekst te analyseren op woord- en zinsniveau. De syntactische structuur en complexiteit van zinnen wordt geanalyseerd om te bepalen hoe leesbaar een stuk tekst is of wat de syntactische stijlkenmerken van een schrijver zijn. Taalkundige annotaties kunnen voor veel verschillende vakgebieden gebruikt worden. In literatuurwetenschap kan de complexiteit van zinnen en het gebruik van namen worden bestudeerd in relatie tot het literaire niveau van een roman. Het automatisch herkennen van namen en relaties in teksten kan voor veel verschillende vakgebieden gebruikt worden.

- *AntCont*³⁴ is een programma voor concordantie en woordanalyse. De frequentie van woorden wordt bijgehouden, en men kan alle voorkomens van een woord in teksten opvragen met de context waarin ze voorkomen (een aantal woorden voor en na het gezochte woord in de tekst).
- *LIWC*³⁵ is een programma waarin woordcategorieën zijn gedefinieerd zoals 'positief emotionele woorden', 'familiegerelateerde woorden', 'causaliteitswoorden' etc. Hiermee kunnen emotionele en cognitieve eigenschappen van teksten gemeten en vergeleken worden.
- *OpenNER*³⁶ is een programma om tekst op vele niveau's te analyseren en annoteren. Individuele woorden worden geannoteerd met informatie over functie in de zin, namen worden herkend en gecategoriseerd (persoon, locatie, organisatie, etc), verwijzingen (woorden als *hij*, *zij*, *die* en *daar*) worden gekoppeld aan hun referenten (de entiteiten in de tekst waar ze naar verwijzen) en zinnen worden gemeten op positief danwel negatief sentiment. Dit programma werkt voor zes verschillende talen.
- *xTas*³⁷ combineert tekstanalyse en -annotatie met *topic modeling* en *document clustering*, e.g. het automatisch identificeren van groepen documenten die vergelijkbare inhoud hebben. Andere mogelijkheden

zijn het detecteren van de taal van een tekst en sentimentanalyse (is een zin positief, negatief of neutraal).

Beeldherkenning en gezichtsanalyse worden gebruikt om bijvoorbeeld gezichtsuitdrukkingen in films te analyseren, maar ook om bepaalde voorwerpen of activiteiten te herkennen, zoals gebouwen, voertuigen, mensen en voetbalwedstrijden.

- *FaceReader*³⁸ analyseert de emotionele uitdrukking in menselijke gezichten in foto's en video's. Hiermee kan men bijvoorbeeld scènes in films en televisieprogramma's analyseren en deze gegevens koppelen aan de emotionele ervaringen van kijkers.
- *MediaMill Semantic Video Search*³⁹ is een set technieken voor het analyseren van en zoeken in collecties van videomateriaal. Aan de hand van positieve en negatieve voorbeelden wordt het systeem getraind om specifieke concepten te detecteren, zoals voetbalwedstrijden, boten, oorlog of specifieke beroemdheden.

Geografische informatiesystemen (GIS) maken het mogelijk om gegevens op een geografische kaart af te beelden. Bekende algemene systemen zijn Google Maps en TomTom, maar er zijn systemen specifiek voor bijvoorbeeld historische gegevens en kaarten.

- *HisGIS*⁴⁰ bevat historische kaarten en gegevens uit o.a. het Kadaster, zoals kaarten met gegevens over grondbezit. Het is ook mogelijk om kaarten uit verschillende perioden over elkaar heen te tonen, zodat bijvoorbeeld kaarten van 1832 en 2013 makkelijk direct en tot in detail vergeleken kunnen worden.
- *QGIS*⁴¹ is een gratis, open source GIS waarmee onderzoekers geo-data kunnen koppelen aan geografische kaarten. Bijvoorbeeld woonplaatsgegevens van historische personen of locaties van gebouwen kunnen via geografische coördinaten in een kaart geplaatst worden.
- *ORBIS*⁴² is een soort Google Maps voor de Romeinse tijd. Met deze tool kunnen wetenschappers een inzicht krijgen in de verschillende routes tussen plaatsen in Europa in de Romeinse tijd, en hoeveel tijd en geld verschillende manieren van vervoer (lopend, rijdend, varend) kosten. Deze getallen in ORBIS zijn gebaseerd op een groot aantal historische documenten waarin vervoersmiddelen, -tijden en -kosten worden genoemd, en worden door een complex wiskundig model berekend.

Linked Open Data is een initiatief om individuele data sets zoveel mogelijk te koppelen aan andere, gerelateerde data sets, zodat een complex

netwerk van semantisch geannoteerde informatie op een gestructureerde manier bevraagd kan worden.

- *PoliMedia SPARQL*⁴³ is een zogenaamd *endpoint*, waar gebruikers gestructureerde zoekvragen kunnen stellen en het systeem de semantische relaties in politieke data gebruikt om precieze antwoorden te geven. De database bevat verslagen van politieke debatten, krantenartikelen en radiofragmenten.
- *Linked Open Data Enhancer* (LODE)⁴⁴ is een platform waar geesteswetenschappers hun gestructureerde data kunnen verkennen, koppelen aan externe gegevensbronnen en verrijken met extra annotaties.

Crowdsourcing is een belangrijke methode voor het verzamelen van data, zoals het beschrijven van afbeeldingen of het transcriberen van handgeschreven tekst. Bij crowdsourcing wordt relatief eenvoudig werk uitgezet op een online platform waar deelnemers taken vrijwillig of tegen een minimale betaling uitvoeren. Er zijn veel succesvolle crowdsourcing-projecten waarbij GW materiaal geanalyseerd wordt (Dunn & Hedges, 2012).

- Het platform *Vele Handen*⁴⁵ heeft collecties van vele instellingen waar duizenden vrijwilligers aan bijdragen. Andere succesvolle projecten zijn *Transcribe Bentham*⁴⁶ en *The Guardian MP Expenses project* (Daniel & Flew, 2010).
- *Amazon Mechanical Turk*⁴⁷ is een algemeen platform voor het opzetten van crowdsource taken waar mensen zich kunnen registreren om taken uit te voeren, meestal tegen een kleine betaling per taak.

Gamification is een gerelateerd concept, waarin taken in een spelvorm worden aangeboden. Gebruikers krijgen punten voor het geleverde werk, waardoor een competitief element ontstaat en gebruikers gemotiveerd raken meer taken te doen, en ze ook sneller en ‘correcter’ uit te voeren. Een bekend probleem is dat in een competitieve vorm, gebruikers ‘correct’ interpreteren als de manier waarop ze punten krijgen en minder belang hechten aan de kwaliteit en het wetenschappelijke doel van het geleverde werk.

- Het eerder genoemde spel *Hooked*⁴⁸ is een voorbeeld van *gamification*, waarmee spelers gegevens genereren waarmee onderzoekers proberen te achterhalen wat muziek zogenaamd *catchy* maakt.

Retrieval

Een zeer belangrijke techniek voor het toegankelijk maken van geesteswetenschappelijk onderzoeksmateriaal is retrieval. Omdat de hoeveelheid beschikbare data zo groot is, zijn slimme zoekmachines essentieel om in die berg materiaal de relevante gegevens te vinden. Er zijn retrievaltechnieken voor een breed assortiment van bronnen (boeken, films, liederen, partituren, kunstwerken, tekeningen, foto's, politieke documenten, etc.) zowel voor het vinden van gehele documenten als voor fragmenten/passages en metadata.

- *Delpher*⁴⁹ is een zoekdienst van de Koninklijke Bibliotheek waarmee ruim 1 miljoen Nederlandse boeken, kranten en tijdschriften doorzocht kunnen worden.
- *Momfer*⁵⁰ is een index van motieven in volksvertellingen, waarmee kan men zoeken naar het gebruik van motieven in specifieke vertellingen of in genres of periodes.
- *Monk*,⁵¹ eerder beschreven hierboven, is een zoekmachine en annotatietool voor handgeschreven bronnen.

4 Digitale Geesteswetenschappen in de Praktijk

DH Gemeenschap

Digitale geesteswetenschappen worden soms bestempeld als een zogenaamde *community of practice*. DH onderzoekers delen methoden en waarden, maar hebben geen gezamenlijk doel. Immers, het object van studie en de bijbehorende onderzoeksvragen van computationele muziekwetenschappen zijn anders dan die van digitale kunstgeschiedenis. Men vindt elkaar vooral in de methoden die gebruikt worden en de gedeelde problematiek bij het bouwen van een degelijk wetenschappelijk fundament waar deze nieuwe methoden op moeten rusten.

Het is interessant dat veel recente initiatieven gericht zijn op de gedeelde problematiek en proberen een zo breed mogelijk geesteswetenschappelijk publiek te benaderen. De achterliggende gedachte is dat DH inclusief is, iedereen is welkom, en dat de gedeelde problematiek het best gezamenlijk opgelost kan worden. Tegelijkertijd levert deze brede aanpak het probleem op dat de groep zeer heterogeen is en zich daardoor zelden kan richten op specifieke problemen voor specifieke vakgebieden. Individuele geesteswetenschappers die willen weten wat

DH te bieden heeft, raken al snel het spoor bijster en missen de specifieke relevantie van DH voor hun eigen onderwerp en vragen.

Er zijn meerdere soorten initiatieven waarmee men DH promoot en de drempel probeert te verlagen voor GW onderzoekers om digitale onderzoeksmethoden te gebruiken.

Ten eerste zijn er initiatieven voor *professionalisering* van GW onderzoekers. Dit gebeurt meestal in de vorm van workshops waarin deelnemers praktische kennis en vaardigheden op kunnen doen in sessies die ingaan op een specifieke tool of methode. Een voorbeeld hiervan is THAT-camp⁵² (The Humanities And Technology camp), een soort unconference waarin de deelnemers op de eerste dag besluiten welke onderwerpen men wil behandelen en een programma met sessies vastlegt, die daarna in zeer informele sfeer plaatsvinden. In een meer vooraf vastgelegde vorm zijn er *crash courses* in digitale tools.^{53,54}

Er zijn ook initiatieven om academisch onderzoek en het bedrijfsleven dichter bij elkaar te brengen middels kleine projecten waarin verkend wordt hoe DH kan bijdragen aan het ontwikkelen van applicaties en diensten voor commerciële doeleinden. Het Joint Center for Digital Humanities,⁵⁵ een samenwerking van de KNAW, UvA en VU, financiert kortlopende *embedded research* projecten, waarin academische onderzoekers samenwerken met bedrijven om DH onderzoek commercieel toepasbaar te maken.⁵⁶

Tenslotte zijn er conferenties zoals Digital Humanities in the BeNeLux (DHBenelux)⁵⁷ met als doelstelling om alle actieve DH onderzoekers en geïnteresseerden bij elkaar te brengen en een lokale gemeenschap te vormen. Hiermee wordt de zichtbaarheid van het lopende onderzoek en DH in het algemeen vergroot.

Complexiteit van geesteswetenschappelijk materiaal

Een van de belangrijkste uitdagingen van DH onderzoek is de grote complexiteit van veel geesteswetenschappelijk materiaal. Die complexiteit uit zich in tenminste drie aspecten:

1. *Incompleteheid*: veel historisch materiaal is incompleet. Door de eeuwen heen is veel materiaal verloren gegaan, maar ook recent digitaal

materiaal zoals websites en tweets verdwijnt. UNESCO heeft zelfs een programma om digitaal erfgoed te beschermen.⁵⁸ De staat van compleetheid bepaalt welke onderzoeksvragen beantwoord kunnen worden, maar vaak is dit onbekend aan het begin van het onderzoek.

2. *Vaagheid van eenheden*: een van de redenen dat digitale analyse veelal op teksten betrekking heeft is dat tekst discreet is, bestaande uit relatief makkelijk te onderscheiden symbolen. Maar auditief en visueel-georiënteerd materiaal is veel lastiger. Wat zijn de eenheden van een schilderij, film of gebeurtenis? Een schilderij bestaat meestal uit penseelstreken, maar hoe kun je die als eenheden in een digitale kopie herkennen en analyseren? Waar houdt de ene penseelstreek op en begint de andere? Hetzelfde geldt voor gebeurtenissen die geen concrete begin- en eindpunten hebben maar wel eindeloos veel niveaus waarop een complexe gebeurtenis in kleinere elementen opgebroken kan worden. Digitale analyse vereist een discretisering van het materiaal, dus er moeten keuzes worden gemaakt voordat de analyse kan plaatsvinden. Maar elke keuze leidt tot imperfecte eenheden.
3. *Niet-tekstueel materiaal en multimodaliteit*: multimodaal materiaal zoals film en video, waarin zowel beelden, geluid, spraak, muziek en tekst gebruikt worden, kunnen op veel manieren geanalyseerd worden, maar alle modaliteiten hangen met elkaar samen. Dit maakt de analyse zeer lastig. Een onderzoek naar de relaties tussen filmbeelden en de ervaring van emoties van kijkers wordt bijvoorbeeld sterk beïnvloed door de aanwezigheid of afwezigheid van geluiden en muziek.

Samenwerken

Een van de kenmerken van digitale geesteswetenschappen is de focus op projectgebaseerd onderzoek en (interdisciplinaire) samenwerking.

GW onderzoekers werken vaak individueel aan hun onderzoek en publiceren hun bevindingen in monografieën. DH onderzoek daarentegen, gebeurt vaak in projectgroepen, waarin onderzoekers uit verschillende disciplines met elkaar samenwerken en ook gezamenlijk publiceren. Wuchty et al. (2007) hebben onderzocht hoe frequent samenwerken en co-auteurschap is in GW, sociale wetenschappen (SW) en exacte wetenschappen (EW). Co-auteurschap van wetenschappelijke publicaties is zeer frequent in EW (ongeveer 80% van alle publicaties

heeft meerdere auteurs) en SW (51%), maar zeldzaam in de GW (minder dan 10%).

Om te zien of digitaal GW-onderzoek tot meer samenwerking leidt heeft Lisa Spiro gekeken naar artikelen in twee tijdschriften tussen 2004 en 2008: een voor traditioneel GW-onderzoek (American Literary History, ALH) en een voor digitaal GW-onderzoek (Literary and Linguistic Computation, LLC). In ALH heeft 2% van alle artikelen meerdere auteurs, in LLC 48% van alle artikelen.⁵⁹

Spiro noemt twee redenen waarom traditioneel GW-onderzoek vaak individueel is:

1. De mythe van het eenzame genie ('myth of the lone genius'): het idee dat men alleen een grote bijdrage kan leveren door zichzelf af te zonderen en niet te laten afleiden door de vele nuttige en niet nuttige meningen van anderen.
2. GW richt zich op bestaande bronnen. De data hoeft niet aan de hand van experimenten gecreëerd te worden.

Er zijn twee andere potentiële oorzaken. Ten eerste is er een mogelijk gebrek aan erkenning voor collaboratief onderzoek. Omdat GW-onderzoek vaak om interpretatie en reflectie draait, is bij een gezamenlijk geschreven publicatie moeilijk te zien wie welke ideeën heeft aangedragen. Wellicht gerelateerd aan de *mythe van het eenzame genie*, is er een cultuur ontstaan waarin onderzoekers worden aangemoedigd zelfstandig hun ideeën uit te werken en te publiceren. Ten tweede leent kwantitatief, empirisch onderzoek zich meer voor samenwerking dan kwalitatief, interpretatief onderzoek. Interpretatie en reflectie zijn inherent subjectief, dus het maakt uit wie het werk doet. Bij kwantitatief onderzoek gaat men er vanuit dat het vastleggen van de methoden en criteria voldoende zijn om het werk objectief uit te voeren, waardoor het niet of nauwelijks uitmaakt wie het uitvoert. Een goed voorbeeld van dit laatste is *crowdsourcing* (zie Sectie 3). Toch is het ook bij kwantitatief onderzoek belangrijk te beseffen dat het nooit puur kwantitatief is. Elke keuze voor de gebruikte methoden, de implementatie daarvan, het selecteren en prepareren van digitaal materiaal en het verwerken van de resultaten bevat kwalitatieve aspecten en interpretatie (zie Sectie 5).

Obstakels

Bij het samenstellen van een projectgroep dient rekening te worden gehouden met de kloof tussen domeinspecifiek jargon en technisch jargon. Een expert op het gebied van kunstgeschiedenis kan aan een programmeur uitleggen wat het onderzoek behelst en wat de te ontwikkelen software moet doen en de programmeur kan vragen stellen over details. Maar de semantische kloof is vaak lastig merkbaar, waardoor er grote kans is op miscommunicatie. De GW-onderzoeker kan niet overzien welke detailbeslissingen nodig zijn en welke basiskennis over kunstgeschiedenis is vereist. De programmeur kan niet inschatten wat de invloed van detailbeslissingen kunnen zijn voor de domeininhoudelijke interpretatie. Hij of zij weet ook niet of beslissingen gebaseerd zijn op eigen, ongetrainde intuïties. Er moet tenminste één persoon zijn die mentaal de brug kan slaan tussen implementatiedetails en specifieke kennis van de discipline.

Deze obstakels wijzen op het probleem van digitale vaardigheden, ook wel *digital literacy* genoemd. Geesteswetenschappers zijn vaak geïnteresseerd in de mogelijkheden van digitale onderzoeksmethoden maar zijn lang niet altijd bereid daarvoor ook zelf de benodigde technische kennis op te doen. Zolang geesteswetenschappers deze methoden niet structureel opnemen in hun onderwijsprogramma's en zichzelf bijscholen, blijft veel DH onderzoek triviaal danwel onbetrouwbaar. Dit probleem ligt ook ten grondslag aan een deel van de kritiek op DH onderzoek (verder besproken in sectie 5) en wijst eigenlijk op een grote behoefte aan wetenschappers die zowel domeininhoudelijke als methodologisch-technische kennis hebben.

Educatie

Er is groeiende aandacht voor DH onderwijs in universiteiten. Meerdere universiteiten in Europa hebben onderwijsprogramma's samengesteld specifiek voor DH. Om te bevorderen dat het wiel niet telkens opnieuw wordt uitgevonden en men van elkaar kan leren wat de belangrijke onderwerpen en technieken zijn, hebben Scagliola en collega's (Scagliola et. al, 2014) een enquête verspreid onder een grote groep digitale geesteswetenschappers om het aanbod van DH-georiënteerd onderwijs in kaart te brengen. De verzamelde gegevens zijn beschikbaar gemaakt in de vorm van een cursusregister.⁶⁰ Het aanbod varieert van kleinschalige minoren, tot volledige bachelor- en masterprogramma's en van discipline-

specifieke programma's tot algemene programma's voor alle GW-disciplines.

5 Kritiek en Weerstand

Er bestaat ook veel kritiek op en weerstand tegen de opkomst van digital humanities.⁶¹ De makkelijke verkrijgbaarheid van digitale onderzoekstools en grote hoeveelheden data, en met name digitale visualisaties, kunnen de indruk wekken dat DH onderzoek weinig moeite kost. Een van de grootste kritiekpunten op demonstraties van digitale tools is dat er een gebrek is aan kritische reflectie en dat niet duidelijk is of deze middelen nieuwe inzichten en kennis opleveren. Als er voorbeelden van toepassingen worden gegeven zijn de resultaten vaak triviaal, en *als* ze verassend zijn, worden er vaak vraagtekens geplaatst bij de methode en/of de data.

Daarnaast maken sommige traditionele geesteswetenschappers zich zorgen dat deze 'nieuwe rage' ten koste gaat van de kritische reflectie en interpretatie die zich door de eeuwen heen vele malen bewezen hebben als uitermate belangrijk voor het vergaren van kennis. Digitale methoden dwingen onderzoekers vaak om een interpretatie of perspectief te kiezen voordat ze hun digitale analysemethoden kunnen toepassen. Dit staat soms haaks op de wens om juist verschillende interpretaties tegelijkertijd te hanteren en tegen elkaar af te zetten. Dit laatste is typerend voor geesteswetenschappelijk onderzoek, en vormt een centraal argument in de bekritisering van DH onderzoek. Wat nodig is in DH onderzoek is het structureel experimenteren met verschillende interpretatie- en implementatiemogelijkheden en de resultaten te vergelijken.

Een deel van de kritiek is terecht, maar ook verraadt de kritiek regelmatig een gebrek aan technische kennis en ervaring met empirisch onderzoek, waarmee de fundering van die kritiek zelf wordt ondermijnd. Hieronder volgt een voorbeeld ter illustratie.

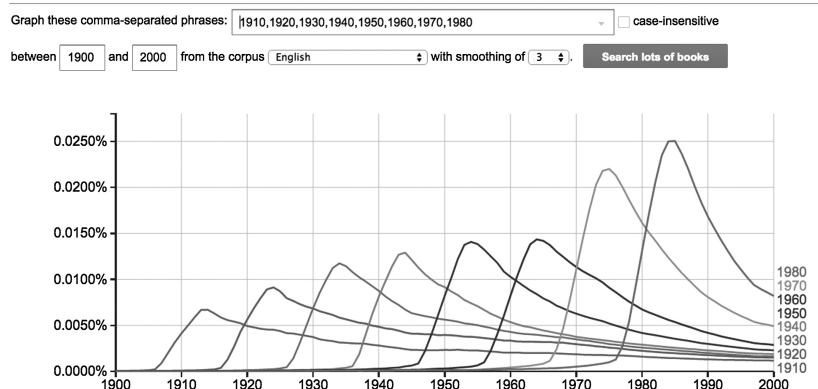
De Google Ngram Viewer

In 2010 lanceerde Google de Ngram Viewer,⁶² een website waarmee gebruikers kunnen visualiseren hoe vaak een woord of frase⁶³ door de

eeuwen heen in boeken is gebruikt. Google heeft immers miljoenen boeken gedigitaliseerd van bibliotheken over de hele wereld.

In een veelbesproken publicatie in *Science* introduceerde Jean-Baptiste Michel en collega's (Michel et al., 2011) het begrip *Culturomics*⁶⁴. De Google Ngram Viewer zou een middel zijn om cultuur op een kwantitatieve manier te analyseren. Door verschillende frases en namen te vergelijken in de Ngram Viewer kunnen allerlei culturele aspecten kwantitatief gevisualiseerd worden. Een van de voorbeelden is de zogenaamde vergeetcurve. Het vergelijken van jaartallen als woorden, dus hoe vaak woorden als '1883', '1945' en '1976' voorkomen, laat een duidelijk en stabiel patroon zien: door een hele reeks jaartallen te vergelijken (1875 tot 1975) zien ze voor oplopende jaartallen steeds hogere pieken, die allemaal terug vallen naar eenzelfde niveau, maar hogere jaartallen zakken sneller terug.

Google books Ngram Viewer



Volgens Michel et al. vergeten we het verleden steeds sneller. Een van de voorbeelden die zij noemen is '1951', dat zelden genoemd wordt tot een paar jaar voor 1951, en piekt in het jaar 1951 (het woord '1951' komt relatief het meest voor in boeken van 1951). Uit de 'Frequently Asked Questions' op de *Culturomics* website blijkt dat er veel kritiek is op deze benadering.⁶⁵ Adam Kirsch (Kirsch, 2014) gebruikt bovenstaand voorbeeld om de opkomst van digital humanities te bekritisieren. De sterke piek van een jaartal als '1951' in de boeken van 1951 is te verklaren door het feit dat in de meeste boeken het publicatiejaar staat, en meer boeken uit 1951 dus het

getal '1951' bevatten dan boeken uit andere jaren. Volgens hem negeren de auteurs dat, waarmee hun conclusies over de vergeetcurve op losse schroeven komt te staan. Het probleem van dit soort gemakzuchtige interpretaties op basis van gebruiksvriendelijke tools wijdt hij aan gebrekkige kennis van de onderzoeksdata en het geesteswetenschappelijke domein. Het lijkt erop dat Kirsch zijn hypothese niet zelf getest heeft. Bovenstaand figuur laat zien dat zijn kritiek onterecht is: van elk jaartal als woord ligt de piek niet in het jaartal zelf, maar een aantal jaren daarna.

Tegelijkertijd bestrijdt hij de noodzaak voor GW-onderzoekers om digitale tools grondig te begrijpen. Immers, om een boek te schrijven hoeft een auteur ook geen gedetailleerde kennis van het boekdrukkerproces te hebben. Deze vergelijking gaat niet op. Het boek als *interface* toont de gehele tekst. Het *interface* van een computerprogramma toont niet de programmacode zelf, maar een buitenkant die de gebruiker afschermt van de interne werking er van. In elk programma dat data analyseert zijn beslissingen genomen door de programmeur die invloed hebben op de uitkomst van de analyse. Zonder inzicht in die beslissingen en de gevolgen daarvan, kan een onderzoeker geen betrouwbare uitspraken doen over de resultaten. Hiervoor is niet per sé toegang nodig tot de software. Door te experimenteren met de mogelijkheden kan vaak indirect afgeleid worden welke keuzes gemaakt zijn.

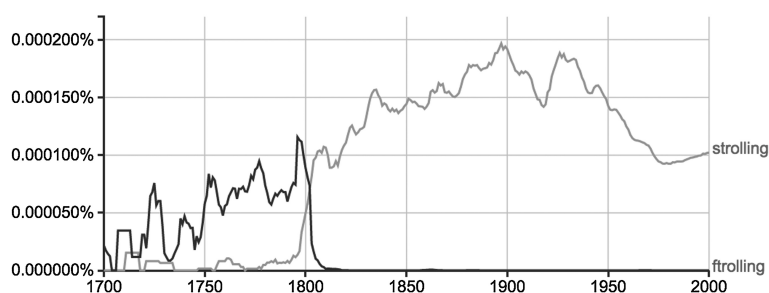
Een relevant voorbeeld komt uit een onderzoek van Joanna Guldi naar varianten van het Engelse werkwoord 'to walk' (Guldi, 2012). Zij gebruikte de Ngram Viewer (versie 2009) om de relatieve frequentie van deze varianten te vergelijken in de periode 1700-2000 en zag rond 1780 een grote toename in frequentie van met name het werkwoord 'strolling'. Ze wijst op de problematiek van het tellen van woorden in gedigitaliseerde teksten, vooral teksten gepubliceerd voor 1900, waar Optical Character Recognition technieken minder accuraat zijn vanwege de slechte kwaliteit van inkt en papier. Ook wordt in deze teksten bijvoorbeeld de letter s vaak voor een f aangezien. Zij ondervangt deze problemen door *close reading*: het zoeken van passages in boeken waarin varianten van 'to walk' voorkomen en deze aandachtig te lezen en interpreteren. Haar conclusie is dat het patroon in de Ngram Viewer correct is.

Echter, een meer recente versie van de Ngram Viewer uit 2012, waarin meer boeken zijn gebruikt en de OCR kwaliteit verbeterd is, laat zien dat het patroon toch niet klopt. Guldi had het gevaar correct geïdentifi-

ceerd: de OCR versie van 2009 had problemen met het herkennen van de letter s in teksten voor 1800. Met een eenvoudig experiment had ze kunnen zien dat dit probleem ook speelde in haar onderzoek. De 2009 versie geeft een sterke opkomst van 'strolling' in 1780 en tegelijkertijd een sterke daling van 'ftrolling'. In de versie van 2012 was dit probleem in ieder geval deels opgelost. In die versie komt 'ftrolling' nauwelijks voor en blijkt de opkomst van 'strolling' veel eerder en geleidelijker te zijn verlopen.

Google books Ngram Viewer

Graph these comma-separated phrases: ☐ case-insen
 between and from the corpus with smoothing of [Search lots of bo](#)



Google books Ngram Viewer

Graph these comma-separated phrases: ☐ case-insensitive
 between and from the corpus with smoothing of [Search lots of books](#)



Kwantitatief onderzoek en interpretatie

Een fundamenteeler kritiekpunt is dat het gebrek aan interpretatie en kritische reflectie in digitaal, kwantitatief onderzoek niet door gemakzucht en gebrekkige kennis komt, maar inherent is aan het soort onderzoek. Stanley Fish schreef een drietal columns voor de New York Times⁶⁶ waarin hij stelt dat digital humanities weinig waardevolle kennis bijdraagt en dat het gebruik van digitale transformaties van teksten, zoals het maken van woordfrequentietabellen en *topic modelling*, elke interpretatie van de resultaten arbitrair maakt. Er wordt immers niet meer naar de betekenis van de tekst als geheel gekeken, maar naar losse elementen uit hun context, waarmee elke interpretatie mogelijk is.

Stephen Ramsay gaat hier tegenin en stelt de vraag wat de interpretatie van de tekst als geheel geldig maakt.⁶⁷ Het kan niet de tekst zelf zijn, want iedere lezer vormt zijn of haar eigen interpretatie, maar waar komt die geldigheid dan vandaan? Een transformatie van een object, digitaal of niet, presenteert het object vanuit een ander perspectief, wat andere interpretaties en reflecties naar voren brengt. De paragrafen van een tekst in willekeurige volgorde lezen (of de shots van een film in omgekeerde volgorde kijken) levert ook andere inzichten op dan van voor naar achteren lezen. Waarom zou het representeren van een tekst als een frequentietabel van woorden geen geldige interpretaties kunnen opleveren? Dit zijn voor geesteswetenschappen wellicht geen gebruikelijke methoden en perspectieven, maar er lijkt geen fundamentele reden om interpretaties naar aanleiding van digitale transformaties uit te sluiten.

6 Toekomst

De toenemende aandacht voor Digital Humanities brengt een spanning tussen hermeneutisch en empirisch GW onderzoek. Sommige GW onderzoekers zien de toenemende rol van empirisch onderzoek als een bedreiging van de trotse traditie van interpretatief onderzoek naar het afwijkende en buitengewone. Het naïeve enthousiasme voor Digital Humanities dat soms gepaard gaat met de puur empirische zoektocht naar patronen wordt niet door iedere geesteswetenschapper met gejuich ontvangen.

Rens Bod (Bod, 2012) roept op om vooral de hermeneutische, interpretatieve traditie te bewaren en te combineren met de kwantitatieve en meer empirische methoden die door digitale technieken makkelijker worden gemaakt. Dit noemt hij ook wel Humanities 3.0, als combinatie van Humanities 1.0 (interpretatief en reflectief) en Humanities 2.0 (kwantitatief en patroonzoekend). Beide benaderingen leveren interessante inzichten op, maar het wordt pas echt vernieuwend en waardevol wanneer digitale methoden verweven worden met kritische reflectie en interpretatie.

Door digitalisering van analoge materialen en de grote hoeveelheid *digital born* data worden de geesteswetenschappen geduwd richting Big Data, wat een paradigmaverschuiving teweeg brengt naar data-gedreven, exploratief onderzoek. Zoals andere wetenschapsgebieden ontkomen ook de geesteswetenschappen niet aan deze *informational turn*. Deze toenemende complexiteit vergt een uitbreiding van het gereedschap waar geesteswetenschappers gebruik van maken. Digital Humanities vervangt niet het traditionele GW onderzoek, maar verrijkt het juist.

Literatuur

- Belter, I. (2014) Informing Physics: Jacob Bekenstein and the Informational Turn in Theoretical Physics. *Physics in Perspective*.
- Bod, R. (2012) Het Einde van de Geesteswetenschappen 1.0. Vossiuspers UvA. http://www.oratiereeks.nl/upload/pdf/PDF-1433Weboratie_Rens_Bod_-_def.pdf.
- Daniel, A. en Flew, T. (2010) The Guardian Reportage of the UK MP Expenses Scandal: a Case Study of Computational Journalism. *Communications Policy & Research Forum*, Sydney.
- Dunn, S. en Hedges, M. (2012) Crowd-Sourcing .Scoping Study: Engaging the Crowd with Humanities Research. <http://crowds.cerch.kcl.ac.uk/wp-content/uploads/2012/12/Crowdsourcing-connected-communities.pdf>.
- Gray, J. (2007) Jim Gray on eScience: A Transformed Scientific Method. In *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.
- Guldi, J. (2012) The History of Walking and the Digital Turn: Stride and Lounge in London, 1808-1851. *Contemporary Issues in Historical Perspective*, University of Chicago Press.

- Hockey, S. (2004) The History of Humanities Computing. *A Companion to Digital Humanities*.
- Honingh, A. en Bod, R. (2012) Clustering and classification of music using interval categories. In *Proceedings of Mathematics and Computation in Music*.
- Jautze, K., Koolen, C., van Cranenburgh, A. en de Jong, H. (2013) From high heels to weed attics: a syntactic investigation of chick lit and literature. *Proceedings of the Second Workshop on Computational Linguistics for Literature*.
- Kestemont, M., Moens, S. en Deploige, J. (2013) Stylometry and the Complex Authorship in Hildegard of Bingen's Oeuvre. *Digital Humanities 2013* (Lincoln, Nebraska, USA), pp. 255-258.
- Kirsch, A. (2014) Technology Is Taking Over English Departments. *New Republic* <http://www.newrepublic.com/article/117428/limits-digital-humanities-adam-kirsch>.
- Meeks, E. en Weingart, S.B. (2012) The Digital Humanities Contribution to Topic Modeling. In *Journal of Digital Humanities*.
- Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176-182.
- Ramsay, S. (2012) Stanley and Me. <http://stephenramsay.us/text/2012/11/08/stanley-and-me/>
- Renear, A.H. (2004) Text Encoding. *A Companion to Digital Humanities*. Blackwell Publishing.
- Scagliola, S., Maas, F. en Stronks, E. (2014) The Teething Troubles of Teaching Digital Humanities: Sharing knowledge and mapping challenges. *DARIAH Workshop DH 2014*, Lausanne.
- Schreibman, S., Siemens, R. en Unsworth, J. (2004) *A Companion to Digital Humanities*. Blackwell Publishing.
- Tangherlini, T. en Leonard, P. (2013) Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research. In *Poetics*, volume 41, nummer 6.
- Traub, M.C., van Ossenbruggen, J., He, J. en Hardman, L. (2014) Measuring the Effectiveness of Gamesourcing Expert Oil Painting Annotations. European Conference on Information Retrieval, Amsterdam, The Netherlands, 2014.
- Weltevrede, E. en Helmond, A. (2012) Where do bloggers blog? Platform transitions within the historical Dutch blogosphere. *First Monday*.

<http://firstmonday.org/ojs/index.php/fm/article/view/3775/3142>.

Wuchty, S., Jones, B.F. en Uzzi, B. (2007) The Increasing Dominance of Teams in the Production of Knowledge. *Science*, vol 316 no. 5827.

Noten

1. De *Day of DH* conferentie doet elk jaar een verzoek aan deelnemers om hun definitie te geven. Zie <http://dayofdh2014.matrix.msu.edu/members/>
2. Melissa Terras, keynote DH Benelux 2014. Zie <http://dhbenelux.org/#Watch-the-keynote-lecture>
3. Zie <http://ashms.uva.nl/research/content/architecture-of-identities-and-a-4d-research-lab/4d-research-lab.html>
4. Zie <http://v-must.net/>
5. Zie <http://cdh.uva.nl/projects-2012-2013/emotions-in-film/emoties-in-film.html>
6. FaceReader is een tool voor het analyseren van emoties in gezichten. Hiervan is een online demo beschikbaar: <http://www.vicarvision.nl/facereader/online-facereader-demo/>
7. Zie <http://www.axiom.humanities.uva.nl/>
8. Zie <https://www.huygens.knaw.nl/circulation-of-knowledge-and-learned-practices-in-the-17th-century-dutch-republic-kccc/>
9. Zie http://ckcc.huygens.knaw.nl/?page_id=1305
10. Zie <http://sappingattention.blogspot.nl/2012/11/reading-digital-sources-case-study-in.html>
11. Zie <http://www.timecapsule.nu/>
12. Zie <http://translantis.wp.hum.uu.nl/>
13. Zie <http://burckhardt.ic.uva.nl/ciage/index.html>
14. De ECARTICO database werd in een project opgezet en nog steeds uitgebreid. Zie <http://www.vondel.humanities.uva.nl/ecartico/>
15. Zie <http://literaryquality.huygens.knaw.nl/>
16. Zie <http://www.hetnationalelezersonderzoek.nl/> voor uitleg en resultaten van dit lezersonderzoek.
17. De documentaire is online beschikbaar, zie <http://mikekestemont.github.io/documentary/>
18. Zie <http://cdh.uva.nl/projects-2012-2013/music-similarity/muzikale-similariteit.html>
19. Zie <http://www.elab-oralculture.nl/cogitch>
20. Zie <http://www.ehumanities.nl/computational-humanities/tunes-tales/>
21. Zie <http://www.cs.uu.nl/research/projects/witchcraft/>
22. Zie <http://socialmediadna.nl/analyse-2/> voor een uitgebreide analyse.
23. Zie <http://www.gabmap.nl/>
24. Zie <http://www.ai.rug.nl/~lambert/Monk-collections-english.html>
25. Zie <http://dariah.eu/>
26. Zie [http://www.nwo.nl/en/research-and-results/programmes/Continuous+Access+To+Cultural+Heritage+\(CATCH\)](http://www.nwo.nl/en/research-and-results/programmes/Continuous+Access+To+Cultural+Heritage+(CATCH))
27. Zie <http://mallet.cs.umass.edu/index.php>
28. Zie <https://radimrehurek.com/gensim/>
29. Zie http://ckcc.huygens.knaw.nl/?page_id=1305
30. Zie <http://mnn.nodegoat.net/viewer>
31. Zie <http://gephi.github.io/>

32. Zie <http://www.vosviewer.com/>
33. Zie <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>
34. Zie <http://www.laurenceanthony.net/software/antconc/>
35. Zie <http://www.liwc.net/>
36. Zie <http://www.opener-project.eu/>
37. Zie <http://xtas.net/>
38. Zie <http://www.facereader-online.com/>
39. Zie <https://ivi.fnwi.uva.nl/isis/mediamill/> (de software is niet direct zelf te gebruiken, maar de site heeft een aantal demonstratievideos)
40. Zie <http://www.hisgis.nl/>
41. Zie <http://www.qgis.org/en/site/index.html>
42. Zie <http://orbis.stanford.edu/>
43. Zie <http://data.polimedia.nl/>
44. Zie <http://www.linkedhumanities.com/>
45. Zie <https://velehanden.nl/>
46. Zie <http://blogs.ucl.ac.uk/transcribe-bentham/>
47. Zie <https://www.mturk.com/mturk/welcome>
48. Zie <http://www.hookedonmusic.org.uk/>
49. Zie <http://www.delpher.nl/>
50. Zie <http://www.momfer.ml/>
51. Zie <http://application02.target.rug.nl/monk/demo.html>
52. Georganiseerd in 2014 (<http://thehague2014.thatcamp.org/>) en 2015 (<http://utrecht.thatcamp.org/about/>)
53. Zie <http://digital.humanities.uva.nl/crash-course-2013-programme/>
54. Zie <http://www.centrefordigitalhumanities.nl/crash-course-2014/>
55. Zie <http://www.centrefordigitalhumanities.nl/>
56. Zie <http://www.centrefordigitalhumanities.nl/embedded-research-projects-2013/embedded-research-projects-2012/> en <http://www.centrefordigitalhumanities.nl/embedded-research-projects-2013/>
57. Zie <http://dhbenelux.org/>
58. Zie http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/charter_preservation_digital_heritage_en.pdf
59. <http://digitalscholarship.wordpress.com/2013/04/10/group-and-method-collaboration-in-the-digital-humanities/>
60. Zie <https://dariah.uni-koeln.de/>
61. Zie o.a. Adam Kirsch (2014) en de opiniestukken van Stanley Fish in de New York Times: <http://opinionator.blogs.nytimes.com/2011/12/26/the-old-order-changeth/>
<http://opinionator.blogs.nytimes.com/2012/01/09/the-digital-humanities-and-the-transcending-of-mortality/>
<http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/>
<http://opinionator.blogs.nytimes.com/2012/01/09/the-digital-humanities-and-the-transcending-of-mortality/>
<http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/>
62. Zie <https://books.google.com/ngrams/>

- 63. Een n-gram is een sequentie van *n* woorden, e.g. *witte huis* is een 2-gram, een combinatie van twee opeenvolgende woorden.
- 64. Zie <http://www.culturomics.org/>
- 65. Zie <http://www.culturomics.org/Resources/faq>
- 66. Voor verwijzingen naar de columns, zie voetnoot 26
- 67. Zie <http://stephenramsay.us/text/2012/11/08/stanley-and-me/>