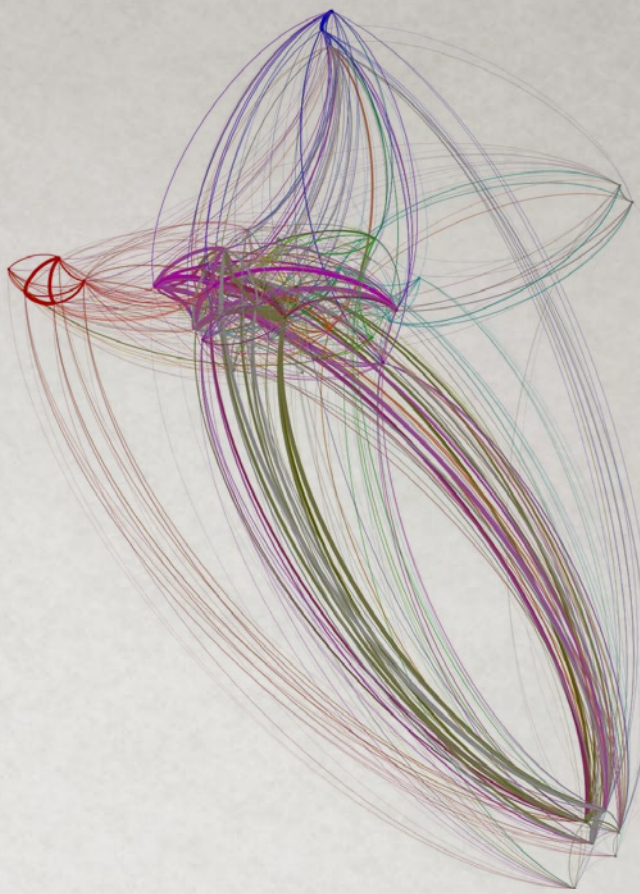


SURFERS VAN DIE TSUNAMI

NAVORSING EN INLIGTINGSTEGNOLOGIE BINNE DIE GEESTESWETENSKAPPE

Burgert A SENEKAL | Susan BROKENSHA

sb

SURFERS
VAN DIE
TSUNAMI

NAVORSING EN INLIGTINGSTEGNOLOGIE BINNE DIE GEESTESWETENSKAPPE

Burgert A Senekal | Susan Brokensha

SUNBONANI SCHOLAR

Surfers van die Tsunami – Navorsing en inligtingstegnologie binne die Geesteswetenskappe

Uitgegee deur Sun Media Bloemfontein (Pty) Ltd.

Druknaam: SunBonani Scholar

Hierdie publikasie is deur die uitgewer aan 'n onafhanklike dubbel-blinde portuurevaluering onderwerp.

ISBN: 978-1-920382-64-3
ISBN: 978-1-920382-65-0
DOI: https://doi.org/10.18820/978-1-920382-65-0

Geset in Minion Pro 12/16
Bandontwerp, bladuitleg en produksie deur Sun Media Bloemfontein

Navorsing en akademiese werke word onder hierdie druknaam in druk en elektroniese formaat uitgegee.

Hierdie publikasie kan bestel word by: media@sunbonani.co.za
Die e-boek is beskikbaar by: https://doi.org/10.18820/978-1-920382-65-0

Inhoud

Lys van figure

Lys van tabelle

Voorwoord

Die tyd toe 'n groot deel van navorsing behels het dat die navorser na 'n biblioteek sou gaan en daar gedrukte artikels sou lees en/of fotostateer, is verby. Hierdie oudmodiese werkswyse is nie alleen onnodig tydrowend nie, maar beteken in die Inligtingsera ook dat daar nie tred gehou kan word met internasionale navorsers nie, omdat daar bloot nie tyd is om so 'n wye verskeidenheid onlangse bronne te raadpleeg as diegene wat wel inligtingstegnologie (IT) inspan nie. As hy nie sy navorsingsmetodes by die 21ste eeu aanpas nie, loop die navorser die gevaar dat hy agterweë kan bly, wat beide sy loopbaangeleenthede en die kwaliteit van studente se onderrig direk kan beïnvloed.

In die 21ste eeu is aanpasbaarheid 'n beslissende faktor vir sukses, ook in die akademiese milieu. Aanpassing by tegnologie is nie opsioneel nie: dit is 'n voorvereiste vir effektiewe werksverrigting. Papp en Alberts (1997:iii) het reeds in 1997 gewaarsku dat ons sukses as individue, families, organisasies, gemeenskappe en samelewings meer as ooit sou afhang van ons vermoë om aan te pas, in byna reële tyd, by die toenemend komplekse en dinamiese situasies wat kenmerkend van die Inligtingsera is. Nietemin bestaan daar soms 'n algemene onwilligheid om hiervolgens aan te pas, indien dit nie selfs op 'n vyandige, of ten minste agterdogtige, houding jeens tegnologie neerkom nie. 'n Ervare navorser het byvoorbeeld by geleentheid teenoor een van die outeurs opgemerk dat die internet oppervlakkig is en dus nie geskik vir wetenskaplike navorsing nie. Só 'n persepsie hou nie rekening met die groot hoeveelheid akademiese publikasies wat aanlyn beskikbaar is nie, en skeer alle bronne wat aanlyn gevind word oor die kam van Wikipedia (wat op sigself ook nie noodwendig onbetroubaar is nie). Tegnologie kom wel met vele probleme, maar desnieteenstaande is dit deel van ons lewe en ons kan dit nie ignoreer nie.

Een van die belangrikste hulpmiddels wat die navorser dus onder die knie moet kry, is die internet. Soos Dolowitz, Buckler en Sweeney (2008:39) opmerk, kan enige navorsingsprojek baat vind by die gebruik van die internet, al is dit bloot om primêre en sekondêre bronne vinniger op te spoor. Bronne wat in digitale formaat beskikbaar is, kan deur middel van die internet opgespoor word, sowel as die fisiese ligging van bronne waarvan daar nie digitale weergawes bestaan nie (byvoorbeeld die meerderheid ouer Afrikaanse boeke). Die gebruik van digitale bronmateriaal word hier sterk aanbeveel, want dit stel die navorser in staat om vinniger en akkurater met groot hoeveelhede inligting om te gaan. Jockers (2013) sluit nie verniet sy boek af met 'n pleidooi dat kopieregprobleme uitgesorteer moet word ten einde die ontleding van digitale

bronmateriaal te bemiddel nie; 'n ontleding soos hy vermag kan geensins sonder digitale bronmateriaal onderneem word nie. Die afwesigheid van digitale boeke kniehalter die digitale geesteswetenskappe in Afrikaans, maar nietemin is daar steeds groot hoeveelhede inligting wat deur middel van die internet opgespoor kan word, byvoorbeeld akademiese publikasies, koerantberigte, en deur sosiale media.

Dit is egter nie voldoende om bloot die internet te raadpleeg nie. Afgesien daarvan dat gewone internetsoektogte slegs deur die oppervlak van die web soek en dus nie die meerderheid inligting kan vind wat aanlyn beskikbaar is nie, moet die navorser in die era van grootdata ('n term wat in hierdie boek breedvoerig behandel word) met meer inligting kan omgaan. Die werklike vraag is nie meer hoe om genoeg inligting te vind, te stoor, te bewaar of selfs te versprei nie, maar hoe om bruikbare inligting uit 'n magdom inligting te herwin (Olcott 2012:95). Hiervoor benodig die navorser nuwe navorsingsmetodes en rekenaarprogrammatuur, 'n nuwe ingesteldheid en ook moontlik 'n paradigmaskuif na die sogenaamde vierde paradigma van die wetenskap (wat dié term behels word ook in hierdie boek bespreek).

Die gebruik van inligtingstegnologie vir navorsingsdoeleindes het beide kwantitatiewe en kwalitatiewe implikasies: nie net kan meer inligting vinniger verwerk word nie, wat tot 'n groter aantal navorsingsuitsette én die nakoming van onderrigverpligtinge kan lei nie, maar dit stel die navorser ook in staat om homself dieper in 'n terrein in te grawe en navorsing van hoër kwaliteit te lewer – omdat die kleiner tydinset in terme van die versameling en ontleding van bronmateriaal die navorser vry laat om meer aandag aan die interpretasie en verwerking van sy onderwerp te wy (Bingham 2010:229). Ook stel inligtingstegnologie die navorser in staat om aansienlik breër na sy onderwerp te kyk as wat tot onlangs toe moontlik was, soos wat Jockers (2013) illustreer met betrekking tot die letterkunde.

Lynch (2008) glo dat die impak van tegnologie op die wetenskap breed beskou moet word. Vir hom behels inligtingstegnologie nie alleen hoëspoedrekenaars en gevorderde rekenaarkommunikasienetwerke nie, maar sluit dit ook gesofistikeerde sensors en ander waarnemings- en eksperimenteringstoestelle wat aan netwerke gekoppel is in, asook sagtewaregedrewe tegnologie wat hoëspoeddatabestuur, -ontleding en -ontginning, en visualisering moontlik maak, sowel as samewerkingsgereedskap en grootskaalse simulasie- en modelleringstelsels. Die gevorderde programmatuur waarna Lynch verwys vorm egter nog nie deel van die hoofstroom binne die geesteswetenskappe in Suid-Afrika of in die buiteland nie, hoewel Borgman (2009:3) noem dat die natuurwetenskappe die geesteswetenskappe vooruit is in die Verenigde State van Amerika en die Verenigde Koninkryk, waar daar onderskeidelik na 'kuberinfrastruktuur' en 'eScience' verwys word. Volgens Borgman bly die toepassing van inligtingstegnologie steeds ontluikend in

die geesteswetenskappe, terwyl eScience reeds die norm binne die natuurwetenskappe geword het. Die geesteswetenskappe hoef wel nie die natuurwetenskappe slaafs na te volg nie, maar nuttige lesse kan geleer word deur die voordele (en beperkinge) van kuberinfrastruktuur en eScience inisiatiewe te bestudeer.

Afgesien van die algemene gebruik van die internet, woordverwerkers en programme soos EndNote en Mendeley vir akademiese doeleindes, bied inligtingstegnologie navorsers die geleentheid om navorsing op 'n nuwe manier te benader. In die buiteland het die term 'digitale geesteswetenskappe' onlangs begin inslag vind.1 Frischer (2009:15) definieer dit as die toepassing van inligtingstegnologie as 'n hulpmiddel om die geesteswetenskappe se basiese take van die behoud, die rekonstruksie, die oordrag, en die interpretasie van die menslike rekord te vervul. Inligtingstegnologie is onder andere al aangewend in die leksikografie (Wooldridge 2004), linguistiek (Hajič 2004), historiografie (Thomas 2004; Schwarte, Haccius, Steenbuck & Steudter 2010), teologie (Kroeze, Matthee & Bothma 2013), en heelwat in die letterkunde (Jockers 2013; Bode 2012; Gottschall 2008; Rommel 2004; Allison, Heuser, Jockers, Moretti & Witmore 2012). Kroeze (2010:918) en Jockers (2013) wys daarop dat rekenaars byvoorbeeld reeds aangewend word om temas en patrone in tekste te identifiseer – iets wat andersins moeilik op 'n groot skaal vermag sou kon word. Tegnologie kan dus in alle fasette waarmee navorsers in die geesteswetenskappe hulself bemoei, aangewend word; trouens, navorsers behóórt dit aan te wend ten einde nie oorweldig te word deur die vloedgolf van data nie.

Verder behoort akademici studente (veral nagraads) te leer hoe om inligtingstegnologie vir akademiese doeleindes te benut om hulle voor te berei vir die werksomgewing. Rekenaarvaardigheid behoort in elke kursus geïntegreer te word, sodat studente daarmee vertroud kan raak en dit met gemak leer benut – as hulle dit nie kan doen nie, stuur ons hulle met 'n agterstand in 'n wêreld in waar hulle internasionaal sal moet meeding met ander navorsers wat wél inligtingstegnologie bemeester het. Voor ons dit egter vir hulle kan leer sal akademici dit self onder die knie moet kry.

Hierdie boek is toegespits op navorsers en doen verslag oor navorsing wat oor die afgelope paar jaar onderneem is om vas te stel hoe inligtingstegnologie aangewend is en kan word vir navorsingsdoeleindes binne die geesteswetenskappe, sowel as watter implikasies die gebruik van inligtingstegnologie vir die geesteswetenskappe inhou in die Inligtingsera. Die beginsels, implikasies, probleme en geleenthede van inligtingstegnologie en die digitale revolusie word teen die agtergrond van grootdata bespreek, en word veral in verband gebring met die geesteswetenskappe in Suid-Afrika.

1 Sien Hockey (2004) vir 'n oorsig oor die ontwikkeling van die digitale geesteswetenskappe.

Erkennings

Die boek is onder andere die resultaat van ’n voortgesette navorsingsprojek wat in 2011 van stapel gestuur is met befondsing van die Erfenisstigting. Hul ruim bydrae het dit moontlik gemaak om rekenaarprogrammatuur te evalueer, waar dié projek sy oorsprong gehad het.

Grootdata en die 'vierde paradigma' van die wetenskap

In 1939 het John Vincent Atanasoff die eerste elektroniese rekenaar ontwikkel. Tydens die Tweede Wêreldoorlog het Alan Turing en John von Neumann op sy werk voortgebou en twee projekte onderskeidelik in Brittanje (by Bletchley Park) en die VSA (by die Universiteit van Pennsylvania in Philadelphia) bedryf. Dié projekte sou die wetenskap en die mensdom onherroeplik verander. By Bletchley Park het Turing en kollegas die rekenaar Colossus ontwikkel, wat onder andere aangewend is om die Duitse Enigmamasjien se kodes te ontsyfer. By die Universiteit van Pennsylvania het Von Neumann, John P. Eckert en John W. Mauchly ENIAC (Electronic Numerical Integrator and Computer) ontwikkel, wat kort ná die oorlog gebruik is om die moontlikheid van die ontwikkeling van 'n waterstofbom te bereken. Alhoewel dié projekte reeds in die dertigs begin is,[2] werk rekenaars vandag nog op dieselfde beginsels wat deur hierdie wetenskaplikes vasgelê is (Dyson 2012:460). Von Bertalanffy (1968:20) het twintig jaar later reeds gelet op watter belangrike impak hierdie ontwikkelings op die wetenskap sou hê, maar dit sou eers in die negentigerjare wees dat rekenaars die wyse waarop die mens met sy wêreld omgaan sou domineer.

Dit is onmoontlik om die invloed van inligtingstegnologie op die mens in die hedendaagse wêreld te oordryf.[3] Inligtingstegnologie het deur middel van die wêreldwye web en sosiale media platforms soos Facebook en Twitter 'n astronomiese impak op sosiale interaksies gehad, wat beide positiewe as negatiewe gevolge inhou waar jongmense se sosiale vaardighede ontwikkel word, maar ook die risiko van bullebakkery inhou (O'Keeffe & Clarke-Pearson 2011). Politiek is onherroeplik verander[4] omdat dit bykans onmoontlik geword het vir regerings om beheer oor inligting uit te oefen, en soos die onlangse Arabiese opstande uitgewys het, kan inligtingstegnologie ingespan word

2 Sien byvoorbeeld Turing (1936).

3 Daar is selfs aanduidings dat die manier hoe mense dink deur inligtingstegnologie verander word (Shroff 2013:6-8).

4 Olcott (2012:82-83) skryf oor hoe die media politieke gedrag in China verander het.

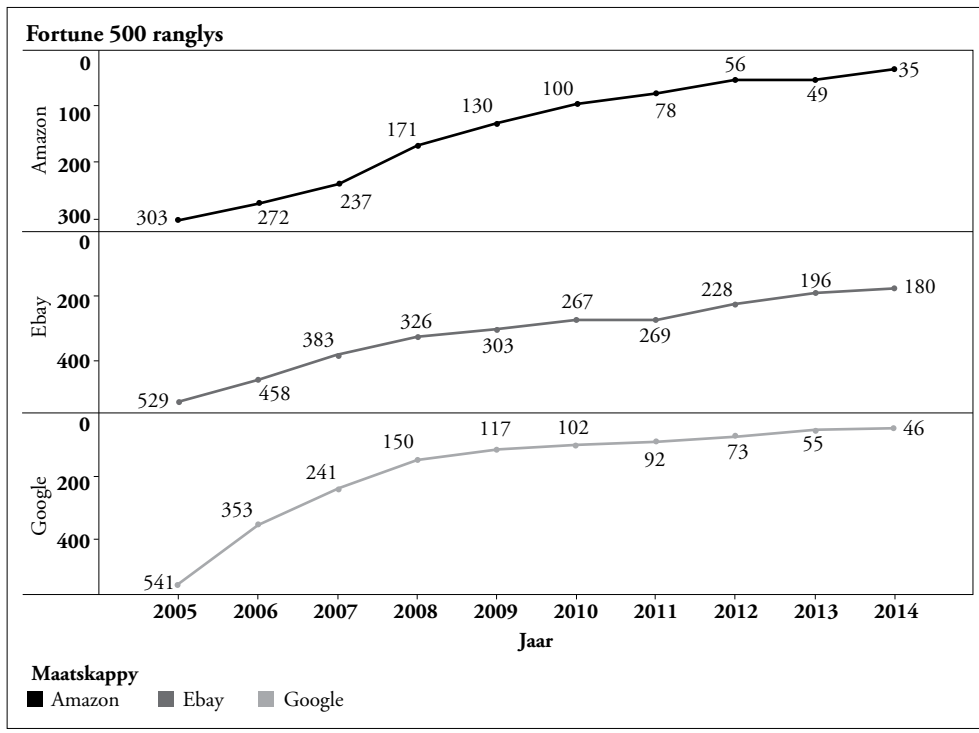
om regerings omver te werp (Kilcullen 2013:179-231). Opvoeding is ook ingrypend verander deur aanlynplatforms soos Blackboard waar studente met kursusmateriaal in die kuberruimte kan omgaan (Chen, Lambert & Guidry 2010; Brokensha 2012), en enige dosent sal kan getuig dat studente gereeld hul inligting vir werkstukke aanlyn kry (ongelukkig ook gereeld sonder 'n bronverwysing). In die publikasie-industrie gebruik uitgewers nie alleen aanlynwinkels om hul boeke te versprei nie, maar word boeke ook in 'n digitale formaat uitgegee – wat nuwe uitdagings en geleenthede vir dié industrie geskep het (Jiang & Katsamakas 2010). Gespreksvoering oor die Afrikaanse letterkunde het tot 'n groot mate ook aanlyn beweeg deur webblaaie soos www.litnet.co.za en www.versindaba.co.za (Senekal 2013), en selfs in die visuele kunste het 'n nuwe medium, generatiewe kuns, onlangs ontstaan wat inligtingstegnologiese fasette soos uitlegalgoritmes as medium aanwend (Lima 2011). Die musiek- en filmbedryf is in 'n nimmereindigende oorlog met die web gewikkel om hul inkomste teen onwettige duplisering en verspreiding te beskerm – iets wat ook die publikasie-industrie direk raak, aangesien digitale weergawes van boeke ook oor die web versprei word (Peitz & Waelbroeck 2006:450). Die wettige digitale verspreiding van musiek het op sigself die industrie onherroeplik verander, onder andere deur geleenthede vir meer kunstenaars te skep om ook hul musiek te verprei (McCubbin 2012). Selfs oorlogvoering word oorheers deur inligtingstegnologie wat militêre intelligensie deurweek (dink byvoorbeeld aan die Predator onbemande lugvaartuig). Inligtingstegnologie kan ook 'n wapen op sigself wees, byvoorbeeld die Stuxnet wurm wat in 2010 na bewering deur die VSA en Israel op Iran losgelaat is om hul kernwapenprogram te ontspoor (Kilcullen 2013:177; Chen & Abu-Nimeh 2011).

In die besigheidswêreld het inligtingstegnologie ook 'n radikale invloed gehad. Inligting is die 'sleutelkommoditeit' (Tinati, Halford, Carr & Pope 2014) in vandag se wêreld, en die inligtingsbedryf het oor die afgelope dekades tot 'n multimiljoendollarbedryf ontwikkel. Van die grootste internasionale maatskappye is betrokke by inligtingstegnologie: natuurlik Microsoft, Oracle, IBM en Apple, asook verskeie ander sagtewaremaatskappye, die telekommunikasiebedryf, ensovoorts. Apple het byvoorbeeld in 2013 'n omset van $170,9 biljoen gehad en daarmee die 5de plek op die Fortune 500-ranglys verower, Hewlett-Packhard was 17de met 'n jaarlikse omset van $112,3 biljoen, IBM 23ste met 'n omset van $99,7 biljoen, Microsoft 34ste met 'n omset van $77,8 biljoen, terwyl Oracle in die 82ste plek was, met 'n omset van $37,1 biljoen (CNN Money 2014). In Tabel 1 volg 'n lys van die voorste internetmaatskappye, sowel as hul jaarlikse omset en rangorde op die Fortune 500.

Tabel 1. 'n Lys van die top internetmaatskappye

Rangorde	Maatskappy	Omset (biljoen $)
35	Amazon.com	74,5
46	Google	60,6
180	eBay	16,1
246	Liberty Interactive	11,3
341	Facebook	7,9
383	Priceline.com	6,8
515	Expedia, Inc.	4,8
522	Yahoo!	4,7

Die werklike interessante faset van dié statistiek lê egter nie in dié maatskappye se rangorde in 'n gegewe jaar nie, maar in hoe hul rangorde verander het. Die grafiek in Figuur 1 stel voor hoe die top drie maatskappye – Amazon, Google, en eBay – se rangorde op die Fortune 500 oor die afgelope dekade verander het.



Figuur 1. Internetmaatskappye se rangordes oor die afgelope dekade

Al drie skuif dus elke jaar met die ranglys op, wat 'n aanduiding is daarvan dat al drie elke jaar al hoe beter vaar in vergelyking met ander maatskappye.

Soos vroeër genoem ressorteer die reusagtige telekommunikasiebedryf ook onder inligtingstegnologie: AT&T is byvoorbeeld die 11de grootste maatskappy ter wêreld, met 'n jaarlikse omset van $128,8 biljoen. Daarbenewens sou 'n mens kon aanvoer dat maatskappye soos LG en Samsung – wat apparatuur soos slimfone vervaardig en wat net sulke astronomiese omsette het – ook betrokke by hierdie bedryf is. Selfs maatskappye wat nie direk by inligtingstegnologie betrokke is nie, soos Walmart of Exxon Mobil (tans die grootste maatskappye ter wêreld in terme van omset), steun op sogenaamde grootdata om produktiwiteit te bevorder en 'n voorsprong bo hul mededingers te behaal en te behou. Trouens, 'n mens sou kon sê dat data en inligting die ruggraat van enige groot en suksesvolle internasionale onderneming vorm, en daar bestaan sterk bewyse (Provost & Fawcett 2013:58; McAfee & Brynjolfsson 2012:64) dat grootdatametodes besighede se werksverrigting verbeter. McAfee en Brynjolfsson (2012:67) stel dit eksplisiet: datagedrewe besluite is gewoonlik beter besluite.

Die verskynsel van grootdata word gereeld na verwys as die 'vierde paradigma'[5] van die wetenskap (Park & Leydesdorff 2013:757; Abreu & Acker 2013:549; Hitzler & Janowicz 2013:233; Kitchin 2014:3), alhoewel die intellektuele geskiedenis van die konsep na die einde van die 19de eeu terugstrek (Faltesek 2013; Barnes & Wilson 2014). Grootdata is die direkte gevolg van die digitale revolusie en is die kern van die Inligtingsera (McNeely & Hahm 2014:304) – astronomiese hoeveelhede data word deur bykans elke maatskappy en elke mens gegenereer (byvoorbeeld deur e-poskommunikasie, aanlyninteraksies, en deur selfone). Soos McAfee en Brynjolfsson (2012:63) skryf is ons elkeen 'n wandelende datagenereerder. Craig en Ludloff (2011:4) merk ons werk aanlyn, ons kuier aanlyn, ons volg nuus en ons gunsteling programme aanlyn, ons dien belasting aanlyn in, ons doen ons banksake aanlyn, ons kan selfs dobbel of seksuele belange aanlyn nastreef, en alles wat ons doen laat 'n digitale voetspoor wat onder grootdata ressorteer. Die vraag is dan hoe om hierdie groot volumes data te berg en te ontgin om tot 'n beter begrip van die wêreld en – wat veral van belang is in die geesteswetenskappe – van die mens te kom, asook wat die impak hiervan op die wetenskap self is en kan wees.

Park en Leydesdorff (2013:756) skryf dat grootdata 'n prioriteit geword het in die akademie, regerings en industrieë, en die geld wat in grootdata belê word, is so reusagtig soos die data self: Wes-Europa belê $2,49 per gigagreep in die bestuur van

5 Die eerste paradigma was die empiriese wetenskap, die tweede die teoretiese wetenskap, en die derde rekenaargedrewe wetenskap (Chen & Zhang 2014:315).

grootdata, gevolg deur die VSA ($1,77), China ($1,31), en Indië ($0,87) (Park & Leydesdorff 2013:756-757). Op 29 Maart 2012 het die Amerikaanse regering hul Big Data Research and Development Initiative bekendgestel, wat verskillende agentskappe betrek in die ontwikkeling van infrastruktuur om grootdata te stoor, te bewaar, te bestuur en te ontleed (Lazar 2012:47; Chen, Mao & Liu 2014:175). Die Amerikaanse regeringsorganisasie DARPA (Defense Advanced Research Projects Agency) – wat verantwoordelik was vir die skepping van die internet – is een van die organisasies wat by die tegnologiese ontwikkeling van grootdata betrokke is, onder andere deur hul program genaamd ADAMS (Anomaly Detection at Multiple Scales). Die VSA se Departement van Verdediging is ook deur 'n program genaamd MAPD (Mathematics for the Analysis of Petascale Data) by grootdata betrokke (Lazar 2012:48). Dit is dan ook die ontwikkelde lande se intelligensiedienste wat veral gebruik maak van grootdata. In die VSA het die NSA (National Security Agency) die PRISM-program (Planning Tool for Resource Integration, Synchronization and Management), terwyl die VK die Tempora-program het (Lyon 2014:2). Beide hierdie inisiatiewe versamel en ontleed groot hoeveelhede data vir intelligensiedoeleindes. Selfs die Verenigde Nasies – wat, soos Davenport (2014:17) mens herinner, nie bekendstaan vir innovering nie – het 'n grootdataprogram genaamd HunchWorks.

In die omgewing van grootdata kom kwessies soos privaatheid en etiek op die voorgrond (McNeely & Hahm 2014:308; Agrawal, Bernstein, Bertino, Davidson & Dayal 2011:10-11). Die NSA stoor na bewering 1,7 biljoen e-posse, telefoonoproepe en ander kommunikasies elke dag (Mayer-Schönberger & Cukier 2013:156), en die voormalige CIA-agent, Edward Snowden, het in Junie 2013 beweer dat hy en ander agente metadata van 3 biljoen telefoonoproepe en interaksies wat deur Facebook, Google, Apple, en ander maatskappye aangeteken is, onderskep het (Van Dijck 2014:197). Maatskappye wat betrokke is in die inligtingsindustrie werk nou saam met sekuriteitsagentskappe (veral in die VSA en Europa) en deel metadata van mense se bedrywighede met dié agentskappe (Lyon 2014). Dit is juis hierdie vennootskap tussen industrie en regering wat kommer wek; uiteraard kan regerings grootdata gebruik om op die bevolking te spioeneer. Inligting wat op sosiale media geplaas word, word ook deurlopend deur intelligensie-agentskappe gemonitor: Leigh van Bryan en Emily Bunting is byvoorbeeld in Januarie 2012 verhoed om die VSA in te gaan na hulle getweet het: "free this week for a quick gossip/prep before I go and destroy America" (Omand, Bartlett & Miller 2012:812). Afgesien van die belangrikheid van konteks in hierdie misverstand, dui hierdie geval ook daarop dat intelligensiedienste wel sosiale media

monitor, en dat regerings as gevolg van grootdata breër insae in mense se aktiwiteite het as ooit tevore (Craig en Ludloff (2011) bespreek in detail hoe hierdie data bekom word).

Die skending van privaatheid het egter ook voordele; slim elektrisiteitsmeters in die VSA en Europa kan elektrisiteitsverbruik monitor en binnenshuise daggaplantasies opspoor (Mayer-Schönberger & Cukier 2013:152-153), terwyl grootdata ook aangewend word om gemeenskappe voorkomend te polisieer, besluite rakende parool vir gevangenes te neem, en toekomstige terroriste te identifiseer (Mayer-Schönberger & Cukier 2013:158-159). Die terroristegroep Al-Shabaab gebruik Twitter om aanvalle te koördineer, Somaliese seerowers gebruik blogs, Twitter en Facebook, Al-Kaïda het destyds 'n webblad genaamd www.alneda.com gehad, en 12 300 oortredings is in 2011 direk aan Facebook gekoppel (Omand, Bartlett & Miller 2012:803-804). Juis omdat sosiale media ook gebruik word vir oortredings is dit nodig vir intelligensie-agentskappe om hierdie datastrome te monitor.

Volgens Mayer-Schönberger en Cukier (2013:160-161) skep grootdata in werklikheid die geleentheid om te ontsnap van stereotipering en groepsidentiteite: 'n enkellopende man met 'n Arabiese naam en 'n eenrigting eersteklasvlug kan moontlik nie meer uitgesonder word as 'n sekuriteitsrisiko nie. Grootdata kan ook help om skuldiges vas te trek deur byvoorbeeld selfoondata te gebruik om te bewys dat 'n verdagte op 'n misdaadtoneel was, en die ander kant van so 'n argument is natuurlik dat dieselfde data die onskuldiges se onskuld kan bewys deur 'n waterdigte alibi te verskaf (Andrejevic & Gates 2014:187-188). Omand, Bartlett, en Miller (2012) skryf verder oor die implikasies en geleenthede van grootdata met spesifieke verwysing na sosiale media, en stel voor (2012:822) dat 'n nuwe akademiese studieveld, sosiale mediastudies, gestig word om metodes te ontwikkel om sosiale media te ontgin.

Die opkoms van grootdata het belangrike implikasies vir elektroniese onderrig en navorsing hierin. Daar is talle voordele van die gebruik van data-ontginning, wat die vermoë om nuttige inligting oor duisende studente in 'n spesifieke aanlyn konteks in te samel, die verbetering van aanwysingsontwerp, en die identifisering van korrelasies tussen studente se akademiese prestasie en die digitale leeromgewing insluit. Data-ontginning op 'n enorme skaal is egter nie sonder probleme nie, en een omstrede kwessie in die gebied van elektroniese onderrig hou verband met etiese oorwegings. Sommige navorsers en opvoeders is onder andere bekommerd oor die inbraak wat die grootdatarevolusie op individue se privaatheid maak (Polonetsky & Tene 2014:29). Wat dié kommer vererger is die vervaging van die onderskeid tussen wat private ruimtes uitmaak en wat openbare forums veronderstel (Bolander & Locher 2014:17).

Dink byvoorbeeld aan 'n scenario waarin 'n dosent ontledings onderneem van sy/haar studente se diskoers soos gevind op Facebook. Sommige navorsers beweer dat die data wat ingesamel word in die publieke domein is, maar wat 'n mens in gedagte moet hou, is dat sulke digitale ruimtes 'n mengelmoes van openbare en private elemente kan wees. Dit is in die publieke domein in die sin dat dit deur 'n groot en anonieme gehoor gelees kan word, terwyl daar terselfdertyd onderwerpe bespreek word wat ons gewoonlik as 'privaat' ag en taal gebruik word wat verband hou met informele en private gesprekke (Landert & Jucker 2011:1423).

Dit is soms onvoldoende om die betrokke individue te beskerm deur die data te 'skrop,' met ander woorde deur persoonlike inligting te verwyder wat 'n uitspraak aan 'n spesifieke persoon koppel. Zimmer (2010:313) herinner aan 'n geval in 2008 waar VSA-gebaseerde navorsers tersiêre studente se Facebook rekeninge versamel het en 'n aantal stappe noukeurig gevolg het in 'n poging om die anonimiteit van die studente en die betrokke instelling te verseker. Ten spyte van hierdie navorsers se pogings – wat die verkryging van etiese klaring van die gegewe instelling se etiese komitee ingesluit het – is die bron van die data vinnig nagespeur en as Harvard College geïdentifiseer. In die nadraai van dié onthulling was die navorsers verplig om hul datastel te onttrek, en het hulle ook onder skerp kritiek deurgeloop. Sonder veroordeling van die navorsers wat betrokke was by die 2008-studie, stel Zimmer (2010:323) voor dat navorsers drie stappe neem om oortredings ten opsigte van privaatheid te voorkom. Dit kom neer op begrip en konseptualisering van wat openbare en private digitale ruimtes is, opleiding rakende die komplekse aard van sosiale media, en die versekering dat kursusse in navorsingsmetodologie die erkenning van die risiko's betrokke by die data-ontleding van aanlyn ruimtes insluit.

'n Addisionele etiese oorweging is die probleem van wat Sara Briggs (2014:4) 'misleiding deur getalle' noem. Briggs (2014:4-5) let op die geval van 'n bekende opvoedkundige sielkundige, Cyril Burt, wat in 1976 postuum daarvan beskuldig is dat hy groot hoeveelhede data in sy ondersoek van tweelinge en die aangebore/aangeleerde-debat vervals het. Daar is soveel teenstanders van Burt as wat daar voorstanders is,[6] maar ongeag sy skuld al dan nie, moet wetenskaplikes wat gebruikmaak van grootdata so deursigtig as moontlik te werk gaan.[7]

6 Ronald Fletcher (2013) verdedig Cyril Burt in *Science, ideology, and the media: The Cyril Burt scandal*.

7 In die hoofstuk oor NVivo ondersoek ons 'n paar van die beginsels van deursigtigheid waaraan navorsers in die geesteswetenskappe kan voldoen om eties en eerlik met die navorsingsproses om te gaan.

Bogenoemde kwessies het beduidende implikasies vir die wetenskap oor die algemeen, en binne die akademie het grootdata só belangrik geword dat joernale oor die afgelope dekade en ’n half ontstaan het wat spesifiek op hierdie terrein fokus, insluitend *Data Science Journal*, *Journal of Data Science*, *EPJ Data Science*, *Giga Science*, *Journal of Big Data*, *Big Data*, *Big Data Research* en *Big Data & Society.* Grootdata lei tans tot ’n hewige debat binne die wetenskap, en daarom fokus hierdie boek op die beweerde implikasies, die probleme en die oplossings van hierdie benadering, met spesifieke verwysing na die geesteswetenskappe.

Die boek is soos volg gestruktureer: Eerstens word ondersoek ingestel na wat grootdata behels, en ’n poging word aangewend om agtergrond te verskaf, sowel as om ’n werksdefinisie saam te stel met inagname van die genuanseerdheid van die konsep binne verskillende velde. Hierna word in gesprek getree met die bewerings rondom grootdata en die moontlike implikasies wat dit vir die wetenskap inhou. Die eerste hoofstukke verskaf dus ’n kontekstualisering, waarna ondersoek ingestel word na hoe die geesteswetenskappe op so ’n wyse met grootdata kan omgaan dat ’n middeweg gevind word tussen die tradisionele wetenskap en radikale grootdatabenaderings. Hier word ’n agtergrond van die wêreldwye web en digitale bronne van data in die Inligtingsera verskaf. Hierop volg ’n bespreking van ’n herposisionering van kwalitatiewe navorsingsmetodes deur middel van rekenaargesteunde ontledings, en dié hoofstuk handel hoofsaaklik oor die gebruik van NVivo vir navorsingsdoeleindes. ’n Volgende hoofstuk bespreek netwerkontleding, hoofsaaklik vanuit ’n visualiseringsoogpunt, met spesifieke verwysing na die geesteswetenskappe. Laastens is daar ’n hoofstuk wat ’n oorsig bied oor die tegnologiese hulpmiddels wat veral met grootdata geassosieer word.

Hoofstuk 1

'n Omskrywing van grootdata

Grootdata is moeilik om te definieer, 'n bewegende teiken waarvan die definisie afhang van die konteks waarbinne die term aangewend word, asook die tegnologie wat beskikbaar is (Schöf 2013:6; Hitzler & Janowicz 2013:233). Shiri (2014:16-18) gee 'n oorsig van verskeie definisies van grootdata, wat veral daarop dui dat grootdata nie alleen 'groot' is omdat dit verbysterende volumes beslaan nie. Franks (2012:4) skryf dat daar geen konsensus bestaan oor hoe om die konsep te definieer nie, maar dat daar altyd 'n aantal gemene delers is wat gebruik word om die konsep mee te omskryf. Dié omskrywing word gewoonlik gedoen in navolging van Doug Laney (2001) se onderskeiding van die drie v's: *volume*, *velocity* (snelheid) en *variety* (verskeidenheid). In sy inleiding tot die joernaal *Big Data*, skryf die redakteur, Edd Dumbill, in soortgelyke terme oor grootdata (2013), soos ook Madden (2012), Olcott (2012), Schöf (2013), Hendler (2013), Syed, Gillela en Venugopal (2013), Chen, Mao en Liu (2014), Chen en Zhang (2014) en ander. Die huidige hoofstuk omskryf grootdata in navolging van hierdie outeurs.

1.1 Volume

In die eerste plek dui die term 'grootdata' natuurlik daarop dat groot hoeveelhede data hier ter sprake is, wat 'n inligtingsoorlading veroorsaak. So 'n inligtingsoorlading word deur Bawden en Robinson (2009:182) gedefinieer as 'n toedrag van sake waar 'n individu se doeltreffendheid in die gebruik van inligting in hul werk bemoeilik word deur die hoeveelheid relevante en potensieel bruikbare inligting wat beskikbaar is. Inligtingsoorlading is egter nie 'n nuwe konsep nie (Blair 2003; Olcott 2012:238). In 1852 is daar in die jaarlikse verslag van die sekretaris van die Smithsonian Instituut in Washington gekla dat die meer as 20 000 volumes wat op daardie tydstip jaarliks gepubliseer is die wetenskap sou oorweldig, tensy hierdie massa behoorlik gerangskik is en 'n manier gevind kon word om die inhoud daarvan te bepaal (Bawden & Robinson 2009:183). Dit was egter eers gedurende die 1990's dat die skaal van die inligtingsontploffing as gevolg van die digitale revolusie sodanige afmetings aangeneem het dat die opspoor van relevante inligting in groot hoeveelhede data die primêre

probleem geword het; voor die koms van die internet was die probleem eerder om genoeg inligting te vind (ibid.:182).

Die inligtingsontploffing kom daarop neer dat die hedendaagse mens gekonfronteer word met 'n 'data-tsunami' wat dreig om hom te oorweldig. In 2005 het die VSA se National Visualization and Analytics Center (2005:2) gewaarsku dat ons vermoë om data in te samel teen 'n vinniger tempo toeneem as ons vermoë om dit te ontleed.8 Die hoeveelheid data wat digitaal beskikbaar is beslaan reeds verbysterende volumes: globaal word daar geskat dat daar teen 2007 reeds 195 eksagrepe se data digitaal op verskeie stelsels gestoor was (Darvill 2011:5). Teen 2013 is daar geskat dat daar 1 200 eksagrepe van data in die wêreld bestaan het, wat só groot is dat as hierdie data op CD's opgeneem sou word, dit vyf afsonderlike hope sou vorm wat tot by die maan strek (Mayer-Schönberger & Cukier 2013:9).9 Een eksagreep bestaan uit 1 048 576 teragrepe, en een teragreep bestaan natuurlik uit 1 024 gigagrepe (sien tabel 2 hieronder), wat beteken dat hierdie data op 44 548 862 911 (44,5 biljoen) DVD's vasgelê sou moes word. Die mens het teen 2012 elke dag 2,5 eksagrepe se data gegenereer, en hierdie getal verdubbel elke 40 maande (McAfee & Brynjolfsson 2012:62) (Wal-Mart verwerk tans soveel data *per uur*). Om dit in ander terme te stel: Meer data beweeg elke sekonde deur die internet as wat 20 jaar terug op die hele internet beskikbaar was (McAfee & Brynjolfsson 2012:62). Boonop neem die generering van data deurgaans eksponensieel toe, en na beraming genereer die mens tussen 2010 en 2015 meer data as wat in die hele geskiedenis van die mensdom gegenereer is (Shroff 2013:xiv). Roberts (2011:9) skryf dat hoewel dit slegs 'n kwessie van tyd is (sommige kenners glo om en by tien jaar) voor die fundamentele beperkings van fisika die rekenaar en grafiese tegnologie sal inperk, ons besig is om asimptoties nader te beweeg aan die perke van die menslike vermoë om data wat ingesamel word te verwerk.

As gevolg van die verbysterende volumes data wat tans bestaan en gegenereer word, skryf Lazar (2012:48) dat dit belangrik geword het om bekend te word met die taal van grootdata, en let onder andere daarop dat alhoewel 'kilo' in gebruik was sedert 1795, 'mega,' 'giga,' en 'tera' almal eers in die 1960's hul verskyning gemaak het, 'peta' en 'eksa' in 1975, en 'zetta' en 'yotta' in 1991, wat juis dui op hoe die hoeveelheid data oor die dekades gegroei het. In Tabel 2 word die groottes van hierdie terme aangedui.

8 Sien ook Honavar (2014:327).

9 Sien ook Chen, Mao, en Liu (2014:171).

Tabel 2. Datagroottes

Naam	Gelykstaande aan	Grootte	Grootte in grepe
Bis	1 bis		0.25
Greep	8 bisse		1
Kilogreep	1024 grepe	10^3 grepe	1024
Megagreep	1024 kilogrepe	10^6 grepe	1,048,576
Gigagreep	1024 megagrepe	10^9 grepe	1,073,741,824
Teragreep	1024 gigagrepe	10^{12} grepe	1,099,511,627,776
Petagreep	1024 teragrepe	10^{15} grepe	1,125,899,906,842,624
Eksagreep	1024 petagrepe	10^{18} grepe	1,152,921,504,606,846,976
Zettagreep	1024 eksagrepe	10^{21} grepe	1,180,591,620,717,411,303,424
Yottagreep	1024 zettagrepe	10^{24} grepe	1,208,925,819,614,629,174,706,176

Dié groottes verg uiteraard nuwe benaderings indien 'n mens wil sin maak van sy omgewing (Agrawal et al. 2011). Die enigste werkbare manier om met groot datastelle om te gaan is deur die gebruik van tegnologie – beide die oorsaak én die oplossing vir inligtingsoorlading – maar dié omgang verg aanpassings van die navorser, soos wat Pirolli en Card (1999:3) aanvoer. Vanuit dié outeurs se oogpunt is die inligtingsverbruiker soos 'n dier wat op inligting voed en meer effektiewe maniere moet ontwikkel word om inligting in die hande te kry, ten einde nie om te kom van die 'honger' nie.

Die probleem met inligtingsoorlading behels dat daar bloot te veel inligting is vir die mens om in ag te neem, maar ook terselfdertyd te veel inligting wat relevant is; McGuire, Stilborne, McAdams en Hyatt (2000:44) noem tereg dat soektogte op die internet soortgelyk daaraan is om 'n slukkie water uit 'n brandkraan te probeer drink. McGuire et al. se stelling is egter lank terug gemaak; vandag sou 'n mens eerder sê dat 'n soektog op die internet is soos om 'n slukkie water uit 'n tsunami te probeer drink. Neri en Pettoni (2009:35) verwys na die 'moderne paradoks': die beskikbaarheid van 'n groot hoeveelheid inligting lei tot 'n inligtingsoorlading, wat die meeste van die tyd geen bruikbare kennis oplewer nie; soms juis die teendeel, soos Patterson et al. (2001:17) beaam. Die kapasiteit van die menslike brein het volgens Darvill (2011:5) gedurende die afgelope 2,5 miljoen jaar verdubbel, maar dit is hopeloos te stadig om met die data-tsunami tred te hou, aangesien Moore se wet bepaal dat verwerkerspoed

en geheuedigtheid elke agtien maande verdubbel (National Visualization and Analytics Center 2005:25).[10]

Die wetenskap is natuurlik ingebed in die Inligtingsera, en inligtingsoorlading het ook 'n direkte effek op die wetenskap. Soos besighede data en inligting aanwend om produktiwiteit te verhoog en inligting dus 'n 'sleutelkommoditeit' van besigheid geword het, is inligting ook 'n kommoditeit binne die akademiese milieu, waar die ontwikkeling van nuwe kennis in die vorm van publikasies, sowel as die oordrag van kennis deur middel van onderrig, die skering en inslag van die akademikus se beroep is. Young, Ioannidis en Al-Ubaydli (2008:2) skryf dat wetenskaplike inligting 'n produk is wat verhandel word in die mark van vaktydskrifte. Daar is reeds heelwat kritiek uitgespreek teenoor die tendens om van die akademiese publikasie 'n kommoditeit te maak, en Castiel en Sanz-Valero (2007:3042) noem dat terme soos "publicationism" en "productivitis" al gebruik is om na die akademiese publikasiebedryf te verwys – met die meegaande siening van dié tendens as 'n 'siekte'. Nietemin is dit die werklikheid van die akademiese milieu dat publikasie tot bevordering lei – 'publiseer of krepeer' – en as die individuele akademikus wil hoop op bevordering, is publikasie onontbeerlik. Dié klem op die toenemende generering van akademiese publikasies het ook 'n data-tsunami binne die wetenskap tot gevolg gehad. Soos Honavar (2014:327) tereg opmerk het publikasies oor die afgelope dekades in 'n verskeidenheid dissiplines radikaal toegeneem. Hy (2014:326) skryf byvoorbeeld dat 2 700 biomediese portuurgroep-beoordeelde artikels *per dag* op PubMed verskyn, wat natuurlik die gevolg het dat geen wetenskaplike ten volle op hoogte kan bly van alle verwikkelinge in sy veld nie. Simon (1971:40) het reeds daarop gelet dat inligting die aandag van die ontvanger 'verbruik'; 'n rykdom van inligting skep dus 'n armoede van aandag en 'n behoefte om aandag doeltreffend te kan toewys aan die oorvloed van inligtingsbronne wat dit kan verbruik. Die navorser word sodoende verplig om sy aandag te verdeel tussen die wye verskeidenheid publikasies wat binne, sowel as buite, sy veld die lig sien, wat 'n armoede van aandag tot gevolg het. Kortom beteken dit dat die akademiese milieu – as deel van die inligtingsbedryf – die akademikus noop om meer koste-effektiewe benaderings tot inligtingsontleding te soek

10 Verwerkerspoed was 10 Mhz in die tagtigerjare (Loukides 2010:3), teenoor huidige snelhede wat in Ghz gemeet word. Grootdata-ontledings, soos byvoorbeeld met behulp van Apache Hadoop, versprei ontledings oor verskeie verwerkers om berekenings vinniger te kan voltrek as waartoe enige enkele verwerker in staat is, hoofsaaklik aangesien die ontwikkeling van verwerkerspoed self nie kan tred hou met die snelheid waarmee data gegenereer word nie.

ten einde sy werkverpligtinge effektief te kan nakom. Die antwoord hier, soos in die geval met besighede, is om inligtingstegnologie in te span.

Bogenoemde groottes is egter nie 'n absolute beraming van die grootte van data nie. Rousseau (2012) let daarop dat wat as 'groot' geag word afhang van die betrokke navorsingsprojek. Vir sommige projekte mag dit 'n aantal teragrepe beteken, terwyl dit in ander projekte mag dui op petagrepe of selfs eksagrepe se data wat ondersoek word. Ook beïnvloed tegnologie wat as 'groot' geag word; wat vandag 'groot' is, is moontlik oor 'n paar jaar hanteerbaar, en wat vir 'n individuele navorser 'groot' is, is nie 'groot' vir 'n maatskappy soos Google of Amazon nie (Franks 2012:24). Boyd en Crawford (2012:663), Davenport (2014:7) en Kitchin (2014:2) skryf dat die grootte van die data om hierdie rede nie die onderskeidende eienskap van grootdata is nie, maar wel 'n belangrike komponent daarvan uitmaak. Russom (2011:6) let weer daarop dat die grootte nie alleen in grepe gemeet word nie, maar ook in die aantal rekords, dokumente, transaksies of tabelle. Ook hang grootte af van hoe die datastel geberg word: dieselfde data kan byvoorbeeld verskillende groottes beslaan afhangende van die formaat waarin dit gestoor is. As boeke se bladsye as TIFF (Tagged Image File Format) gestoor word, kan dit tot 80 megagrepe per bladsy beslaan (afhangende van die resolusie waarmee dit geskandeer is) (Senekal 2011:55), terwyl 'n onbewerkte teks-dokument (.txt) 'n klein aantal kilogrepe sal beslaan – ten spyte van die feit dat beide weergawes dieselfde inligting sal bevat. Daar is selfs verskille in groottes tussen weergawes van Microsoft Excel of Word dokumente.

Deels omdat die grootte van die data wat ondersoek word 'n relatiewe begrip is, skryf Mayer-Schönberger en Cukier (2013:29) dat dit nie die grootte van die datastel in absolute terme is wat dit as grootdata eien nie, maar die *omvattendheid* van die datastel. In hul siening lê die belangrikste verskil tussen klein- en grootdata daarin dat die *hele* datastel in grootdata-ontledings ondersoek word, terwyl slegs 'n steekproefneming in tradisionele, kleindata-ontledings gedoen word.[11] Tegnologie het dit moontlik gemaak om omvattende datastelle te versamel, en rekenaarprogrammatuur kan vandag omgaan met die hele datastel in plaas van slegs 'n komponent daarvan. In hierdie opsig is Ferrer (2013) se studie van die Kanadese literêre kanon 'n grootdatastudie (soos sy dit noem), aangesien sy na 'n omvattende geheelbeeld kyk, soos ook Senekal (2013; 2014) se studies van die hedendaagse Afrikaanse poësie, en Jockers (2013) se ontleding van 'n hele korpus van meer as 3 000 negentiende-eeuse Engelstalige romans. Enige studie wat

11 Sien ook Loukides (2010:3), Tinati et al. (2014:665), en Fan, Han en Liu (2014:2).

die geheelbeeld ondersoek is dus in hierdie opsig 'n grootdatastudie, al is die data nie 'groot' in grepe nie, en al beslaan dit nie miljoene dokumente of datapunte nie.

Die belangrikheid van die ontleding van omvattende datastelle lê daarin dat die geheel eienskappe vertoon wat nie in die onderdele teenwoordig is nie – 'n konsep wat met die teorie van kompleksiteit skakel (McNeely & Hahm 2014:305) en immers reeds vele male binne die sisteem- en netwerkteorie geopper is ('die geheel is meer as die somtotaal van die onderdele' soos Aristoteles dit gestel het).[12] Volgens Mayer-Schönberger en Cukier (2013:10) skep grootdata die geleentheid om tot nuwe insigte rakende die wêreld te kom omdat nuwe betekenis opgesluit lê in groter datastelle, en om omvattend na verskynsels te kyk. Hulle (2013:11) noem swaartekrag om te verduidelik hoe 'n klein of groot skaal die funksionering van die werklikheid beïnvloed. Terwyl swaartekrag 'n groot invloed op die mens uitoefen, geld dieselfde byvoorbeeld nie vir klein insekte nie; insekte wat op water loop se wêreld word eerder beïnvloed deur kapillêre kragte as swaartekrag. Op dieselfde manier is die betekenis wat opgesluit lê in grootdata volgens die outeurs veel anders as wat tot op hede ontdek kon word.

Schreibman, Siemens en Unsworth (2004:xxvi) voer aan dat inligtingstegnologie 'n navorser in die geesteswetenskappe in staat stel om onder andere verbande tussen tekste, asook patrone, te identifiseer wat hy nie daarsonder sou kon herken nie, en dit gebeur veral omdat datastelle op 'n groter skaal bestudeer kan word. Nie alleen verskaf inligtingstegnologie 'n noodsaaklike manier om groot volumes inligting te hanteer nie, maar skep dit ook die geleentheid om vanuit 'n ander invalshoek na bronmateriaal te kyk. Hoewel inligtingstegnologie in byna enige veld die belofte inhou om die mens toe laat om dieselfde take beter en vinniger te vermag, is die meer fundamentele resultaat hiervan dikwels die vermoë om heeltemal nuwe dinge te kan doen (Besser 2004:558).

Neem byvoorbeeld Christakis en Fowler (2007) se kontroversiële studie van die verspreiding van vetsug: die outeurs het aangetoon dat vetsug tot in die derde graad oordraagbaar is, met ander woorde 'n mens se risiko om gewigsprobleme te ontwikkel vergroot wanneer jou vriende se vriende gewig optel. Só 'n tendens is onmoontlik om met klein datastelle te merk – die hele sosiale netwerk is nodig om te sien hoe sulke indirekte invloede versprei. Dié is by uitstek 'n studie wat nie sonder grootdata gedoen sou kon word nie: die outeurs het die sosiale netwerke van 12 067 mense vanaf 1971 tot 2003 bestudeer om hierdie verspreiding te identifiseer. Die skaal van die ondersoek laat die navorsers dus toe om nuwe dinge te merk (daar word later teruggekeer na dié studie).

12 Sien o.a. Von Bertalanffy (1972:407).

Omvattendheid is ook belangrik vir die klein-wêreld-fenomeen: Milgram (1967) het aangetoon dat mense gemiddeld met 'n klein aantal stappe van mekaar verwyderd is – 'n idee wat in die Inligtingsera in populêre kultuur gerealiseer is deur die Kevin Bacon-, Monica Lewinsky- en Marlon Brando-speletjies – maar sy subjekte kon nie die *kortste* paaie vind nie omdat hulle nie 'n geheelbeeld gehad het nie. Neem byvoorbeeld die hedendaagse Afrikaanse filmakteurnetwerk, met data wat deur een van die outeurs (Senekal) saamgestel is: wat is die kortste pad tussen Naas Botha en Steve Hofmeyr? As 'n mens oor die hele datastel beskik wat aandui wie in watter film gespeel het, kan dit maklik bepaal word: Naas Botha het 'n rol gehad in *As jy sing* saam met Hanna Grobler, wat 'n rol vertolk het in *Platteland* saam met Steve Hofmeyr ('n alternatiewe pad loop deur *100m Leeuloop*, waarin beide Naas Botha en Hanna Grobler rolle vertolk het). Wat van tussen Anna-Mart van der Merwe en Steve Hofmeyr? Anna-Mart van der Merwe het 'n rol vertolk in *Die Ballade van Robbie de Wee* saam met Richard van der Westhuizen, wat 'n rol vertolk het in *Bakgat 3* saam met Steve Hofmeyr. As die hele datastel nie in berekening gebring word nie, kan sulke kortpaaie natuurlik nie uitgewys word nie. In die hele filmakteurnetwerk (met 1 715 akteurs) is akteurs gemiddeld slegs 2,33 stappe verwyderd van mekaar, en op die meeste vier stappe. Dié voorbeeld mag ligsinnig voorkom, maar kortpaaie is belangrik in sosiale netwerke, aangesien dit dui op hoe vinnig idees, gerugte, invloed, inligting, en siektes kan versprei. Om die kortste pad te vind word die hele datastel benodig, en die feit dat komplekse netwerke deur 'n kort gemiddelde pad gekenmerk word kon eers geïdentifiseer word met die koms van die internet en die beskikbaarheid van groot digitale datastelle.[13]

Albert-Lázló Barabási se studies rakende wat hy opwellingheid (*burstiness*)[14] noem illustreer ook die waarde wat in omvattende datastelle opgesluit lê (Barabási 2005b; Goh & Barabási 2008; Oliveira & Barabási 2005; Barabási 2011). Mense se optrede is natuurlik onvoorspelbaar wanneer slegs na individuele gevalle gekyk word, maar wanneer groter datastelle ontleed word, kom patrone uit die verf. Barabási het byvoorbeeld ondersoek ingestel na die patroon waarmee mense op e-posse antwoord, en

13 Sien byvoorbeeld Watts en Strogatz (1998).

14 Opwellingheid toon sterk ooreenkomste met wat Kwapień en Drożdż (2012:220) na verwys as die Josef- en Noag-effekte, wat onderskeidelik verwys na patrone wat herhaal of skielike omwentelings wat die bestaande orde omverwerp (sien ook Jones en Breunig (2007:331)). 'n Mens sou opwellingheid ook in verband kon bring met paradigmaskuiwe in die wetenskap of met die opkoms en veranderinge van genres en literêre bewegings (dink byvoorbeeld aan die Sestigers).

hierna ook gekyk na Albert Einstein en Charles Darwin se korrespondensiepatrone om te bepaal of e-pospatrone beperk is tot elektroniese kommunikasie (Oliveira & Barabási 2005). In laasgenoemde studie is gekyk na die responskoers en -tempo waarmee Darwin 7 591 briewe gestuur en 6 530 ontvang het, sowel as Einstein, wat meer as 14 500 briewe gestuur en 16 200 ontvang het (Oliveira & Barabási 2005:1251). Barabási het deurgaans reëlmatige kommunikasiepatrone gevind, wat byvoorbeeld onafhanklik is van ouderdom, tegnologie of persoonlikheid. Sulke bevindinge kan nie op die mikroskaal waargeneem word nie, maar is juis sigbaar wanneer op makroskaal na mense se handelinge gekyk word (Lansing 2003:185).

Omvattendheid is belangrik in die letterkunde. Dit is lank reeds 'n aanvaarde siening dat literatuur nie in isolasie funksioneer nie, maar as 'n sisteem (sien Senekal 1987). Moretti (2005:4) beklemtoon dat die letterkunde as sisteem funksioneer en daarom nie begryp kan word deur afsonderlike brokkies inligting saam te voeg nie – dit moet as geheel bestudeer word. Jockers (2013) is tot op hede een van die outeurs wat binne die letterkunde skryf wat grootdata optimaal benut het (hy begin ook sy boek met 'n verwysing na die grootdatavoorstaander, Anderson (2008)). Jockers gebruik data-ontginning, algoritmes, statistiese metodes en die visualisering van data meer as enige ander skrywer binne die letterkunde om patrone tussen tekste aan te dui wat sonder twyfel nie daarsonder gedoen sou kon word nie. In *Macroanalysis* ontleed hy byvoorbeeld 'n korpus van 3 346 negentiende-eeuse romans. Soos Moretti voer Jockers (2013:32) ook aan dat dit belangrik is om literatuur nie alleen te bestudeer as 'n (klein) aantal verteenwoordigende tekste nie, maar as 'n 'ekosisteem' van tekste wat onderling saamhang. Wanneer literatuur op hierdie skaal bestudeer word, kom interessante patrone uit die verf: Amerikaanse tekste in hierdie tydperk gebruik byvoorbeeld die bepaalde lidwoord meer gereeld as hul Britse eweknieë, maar die Amerikaanse en Britse gebruiksfrekwensies is gekorreleer, wat volgens Jockers 'n duidelike interaksie tussen dié literêre sisteme aandui. Vroueskrywers (uit 'n korpus van 1 363 romans) skryf ook gewoonlik oor onderwerpe wat met tradisionele geslagsrolle saamhang, byvoorbeeld kinders, emosies en klere, terwyl manlike skrywers (uit 'n korpus van 1 753 romans) veral skryf oor wapens en oorlogvoering (Jockers 2013:136 e.v.). Sulke literêre feite kom slegs tot die aandag van die navorser wanneer 'n hele korpus tekste ontleed word.

1.2 Snelheid

Wat die snelheid aanbetref, word daar in ontledings van grootdata gewoonlik verwys na data wat deurlopend op 'n groot skaal gegenereer word, byvoorbeeld deur 'n groot

aantal sensors, of deur sosiale media. Hedendaagse tegnologie het dit moontlik gemaak om bykans alles te monitor, van die gebruik van voertuie deur koeriermaatskappye (Davenport 2014:178), tot die werksverrigting van parte in 'n Boeing (Csermely 2006:91), tot persoonlike roetines en gewoontes. Davenport (2014:12) noem byvoorbeeld die Nike en iPod kombinasie wat in 2006 geloods is en dit moontlik maak om 'n mens se oefenroetines te monitor. John Deere beplan ook om sensors in hul trekkers te installeer (Davenport 2014:47), en daar is planne om sensors in motors te installeer wat soortgelyk aan 'n vliegtuig se vlugopnemer werk om sodoende die oorsake van motorongelukke agterna te identifiseer (Craig & Ludloff 2011:7-8). Die data wat dan deurlopend deur hierdie sensors gegenereer word, moet in reële tyd ontleed word as dit enigsins bruikbaar wil wees vir besigheidsdoeleindes.

Sosiale media genereer natuurlik ook deurlopend massiewe hoeveelhede data. Facebook groei byvoorbeeld met 500 teragrepe per dag (Hendler 2013:18; Kambatla, Kollias, Kumar & Grama 2014:2562), insluitend 2,7 biljoen 'Likes' en 300 miljoen nuwe foto's (Kitchin 2014:2). 'n Totaal van 9 100 tweets word *elke sekonde* op Twitter geplaas (Kambatla et al. 2014:2562). Groot internasionale maatskappye se datavloei is net so astronomies: eBay verwerk daagliks 100 petagrepe se data, terwyl Walmart 2,5 petagrepe se data rakende 1 miljoen transaksies *elke uur* genereer (Kitchin 2014:2; Kambatla et al. 2014:2562). Hierdie data word dan ook in reële tyd ontgin en ontleed (Schöf 2013:5-6; Tinati et al. 2014:665): Google verwerk daagliks 4 biljoen soekresultate (Shroff 2013:5), en soek deur meer as 20 petagrepe se data (Hendler 2013:18), waaronder 3,5 miljoen nuusartikels (Owen, Anil, Dunning & Friedman 2012:5). Vir besighede is die vermoë om groot hoeveelhede data in reële tyd te ontleed van groot waarde, aangesien hulle soekresultate deurlopend kan verbeter (in die geval van Google of Amazon). So ook kan kooppatrone gemonitor word, wat tot 'n meer doeltreffende bemarkingstrategie kan lei (byvoorbeeld eBay of Amazon). Verbruikers se houdings jeens 'n produk kan ook vinnig deur die monitering van sosiale media bepaal word, en oor die algemeen stel die ontginning van data in reële tyd besighede in staat om meer gefokusde bemarking toe te pas waar die individu as 't ware *as* 'n *individu* geteiken word. In 'n grootdata-bemarkingsveldtog ontvang die individu byvoorbeeld nie aanbiedinge wat hom nie raak nie: Amazon en eBay maak voorstelle van produkte waarin hul glo die individu belangstel, en dié voorstelle is tot 'n groot mate sinvol omdat dit spesifiek gemik is op die koper se behoeftes soos bepaal deur data versamel uit hulle internetgebruik en -voorkeure. Indien só 'n bemarkingsveldtog suksesvol wil wees, moet data rakende verbruikers se kooppatrone in reële tyd versamel en ontleed word. Vir militêre intelligensie is die

deurlopende generering en ontleding van data natuurlik ook van onskatbare belang, aangesien dit juis nodig is om sekuriteitsrisiko's so vinnig as moontlik te identifiseer, en hier is die uitdaging ook net so beduidend: onbemande lugvaartuie soos die Predator het teen 2011 soveel videomateriaal gegenereer dat daar daagliks 1 500 uur se video (benewens 1 500 foto's) verwerk moes word (Olcott 2012:105-106).

Wat die belangrikheid van hierdie aspek van grootdata aanbetref is daar egter verskillende sienings, wat hoofsaaklik afhang van die agtergrond waaruit grootdata benader word. Vir Davenport (2014) – wat vanuit 'n besigheidsagtergrond buite die wetenskap skryf – is die snelheid waarmee data gegenereer en ontleed word een van die belangrikste eienskappe van grootdata (tesame met die verskeidenheid wat in die volgende onderafdeling bespreek word). Hy stel voor dat hierdie eienskap die meeste uitdagings en geleenthede bied, byvoorbeeld vir bemarkingsdoeleindes en om kostebesparings te bewerkstellig. Volgens Schöf (2013) – wat vanuit die geesteswetenskappe skryf – is hierdie aspek van grootdata egter van minder belang, en Jockers (2013) fokus ook nie op hierdie aspek van grootdata in sy benadering tot die letterkunde nie. Ook sou 'n mens kon aanvoer dat die geesteswetenskappe oor die algemeen nie só 'n groot klem daarop plaas om data in reële tyd te ontleed nie, anders as besighede en militêre intelligensie. Dit sal 'n navorser min baat om data in reële tyd te kan ontleed en dan sy bevindinge aan 'n joernaal voor te lê wat 'n jaar of twee mag neem om die studie te publiseer. Omdat hierdie aspek van grootdata van minder belang vir die geesteswetenskappe is, word daar nie in die huidige boek op hierdie aspek van grootdata gefokus nie.

1.3 Verskeidenheid

Schöf (2013:4) skryf dat grootdata gewoonlik in verskillende formate bestaan, met ander woorde in 'n verskeidenheid formate geënkodeer is, beide gestruktureerd as ongestruktureerd. Gestruktureerde data is data wat in 'n geordende formaat aangeteken is (Syed, Gillela & Venugopal 2013:2446), byvoorbeeld in 'n Microsoft Excel of Access dokument, of op 'n tradisionele databasis soos MySQL of Inmagic DBText. Onbewerkte teks-dokumente (.txt) word ook as gestruktureerd geag. Semi-gestruktureerde data is XML (Extensible Markup Language), waar die verhouding tussen komponente duidelik volgens 'n skema aangeteken is, maar nie in tabelle en kolomme geberg is nie. Ongestruktureerde data kan gesien word in die meerderheid inligting wat opgesluit lê in dokumente soos Word, PDF, verskeie beelde (onder andere JPG, TIFF, PNG, ensovoorts), klank- en video-opnames, die inhoud van e-posse, en aanlyninteraksies soos Facebook en Twitter boodskappe (Syed, Gillela & Venugopal 2013:2446;

Russom 011:7). Gestruktureerd verwys in hierdie geval na gestruktureerdheid vanuit 'n rekenaar se oogpunt; natuurlik is taal gestruktureerd, en enige teks het uiteraard 'n struktuur, al ag 'n rekenaar dit nie as gestruktureerd nie. Uit die oogpunt van 'n rekenaar is gestruktureerde data dié data wat in tabelle en kolomme geberg is.

Tabel 3 gee die mees algemene tipes dokumente weer, sowel as of dit gestruktureerd of ongestruktureerd is.

Tabel 3. Formate van dokumente

Lêerekstensie	Tipe dokument	Formaat
.accdb	Microsoft Access Database	Gestruktureerd
.avi	Audio Video Interleaved	Ongestruktureerd
.bmp	Bitmap	Ongestruktureerd
.csv	Comma Seperated Value	Gestruktureerd
.doc	Microsoft Word Document	Ongestruktureerd
.docx	Microsoft Word Document (2007 en later)	Ongestruktureerd
.exe	Executable File	Ongestruktureerd
.gif	Graphics Interchange Format	Ongestruktureerd
.jpeg	Joint Photographic Experts Group	Ongestruktureerd
.mp3	Moving Picture Experts Group	Ongestruktureerd
.mpeg	Moving Picture Experts Group	Ongestruktureerd
.pdf	Portable Document Format	Ongestruktureerd
.png	Portable Network Graphics	Ongestruktureerd
.ppt	Microsoft Powerpoint Presentation	Ongestruktureerd
.pptx	Microsoft Powerpoint Presentation (2007 en later)	Ongestruktureerd
.rtf	Rich Text Format	Ongestruktureerd
.tiff	Tagged Image File Format	Ongestruktureerd
.txt	Text	Gestruktureerd
.wav	Waveform Audio File Format	Ongestruktureerd
.wmv	Windows Media File	Ongestruktureerd
.wpd	Word Perfect Document	Ongestruktureerd
.xls	Microsoft Excel Spreadsheet	Gestruktureerd

Grootdata word gekenmerk daardeur dat dit in 'n verskeidenheid formate aangeteken is, insluitend ongestruktureerde formate, wat ontleding bemoeilik. Grootdata is data wat "wild" (Loukides 2010:3; Tinati et al. 2014:665) voorkom, met ander woorde data wat nog nie georden is nie. Om sulke datastelle te ontleed word nuwe rekenaarprogrammatuur benodig wat die gestruktureerde/ongestruktureerde hindernis kan oorbrug.

Lues en Lategan (2006:6) let daarop dat die navorser tydens die verwerkingsfase van die navorsingproses rou data in meer bruikbare data omskep deur dit, onder andere, te sorteer en groepeer. Dié fase word van groter belang in 'n grootdata-omgewing: Agrawal et al. (2011:4-5) skryf dat baie data eers in 'n gestruktureerde formaat omgeskakel moet word voor dit ontleed kan word, wat 'n verdere probleem skep omdat dit die verwerkingsfase van die navorsingsproses vergroot. Russom (2011:17) let daarop dat gestruktureerde data steeds grootdata-ontledings in die besigheidsektor domineer. Uit Davenport (2014:19, 100) se navorsing het dit geblyk dat tot 80% van tyd spandeer word om data in die regte formaat te kry sodat dit as gestruktureerde data ontleed kan word, en dit geld beide vir besigheid as regeringsorganisasies.[15] Tableau, wat hieronder as 'n grootdata-ontledingsprogram genoem word, kan byvoorbeeld slegs gestruktureerde data hanteer, en dieselfde geld vir die meerderheid netwerkontledingsprogrammatuur (behalwe Palantir en Starlight VIS). Die navorser wat dan hierdie programmatuur wil aanwend vir ontledingsdoeleindes word genoop om eers ongestruktureerde data in 'n gestruktureerde formaat om te skakel, en dít is veral belangrik binne die geesteswetenskappe, waar min data in 'n gestruktureerde formaat beskikbaar is. Die datastel rakende die hedendaagse Afrikaanse filmindustrie waarna vroeër verwys is, is juis deur so 'n omskakelingsproses, aangesien die data aanvanklik as video's bestaan het (die krediete van die films self), maar in 'n Microsoft Excel-formaat omgeskakel moes word om ontleed te kan word.

Indien die navorser oor rekenaarprogrammeringsvaardighede beskik, kan ongestruktureerde data makliker in 'n gestruktureerde formaat omgeskakel word deur byvoorbeeld LDA (Latent Dirichlet Allocation) (Blei, Ng & Jordan 2003; Blei, Griffiths, Jordan & Tenenbaum 2004; Griffiths & Steyvers 2004) of MALLET (MAchine Learning for LanguagE Toolkit) (McCallum 2002), maar só 'n omskakeling verg gewoonlik rekenaarvaardighede waaroor die navorser nie altyd beskik nie (Jockers is 'n uitsondering). Soos hieronder in meer besonderhede bespreek sal word, noop die formaat

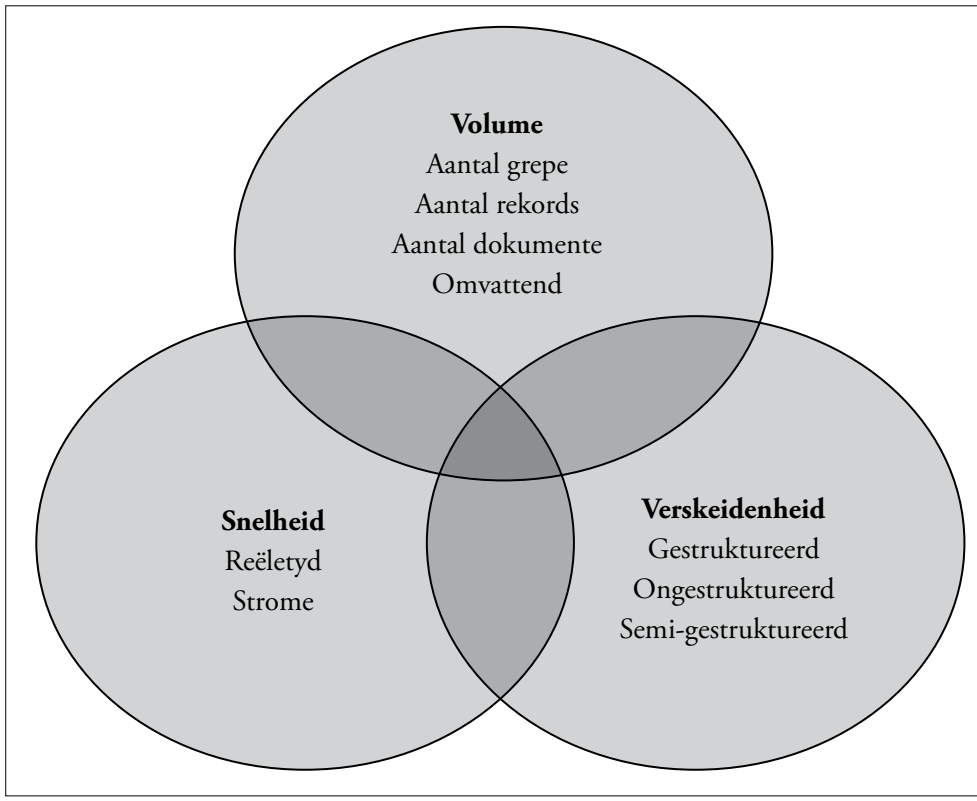
15 Sien ook Franks (2012:16).

van die dokumente die navorser om interdissiplinêre samewerkingsooreenkomste te sluit met rekenaarwetenskaplikes, óf om nuwe rekenaarprogrammatuur aan te wend en rekenaarvaardighede te ontwikkel. Let daarop dat hierdie probleem selfs nog nie volkome binne besigheid of militêre intelligensie aangespreek is nie, ten spyte van die feit dat dié terreine die toonaangewende velde in die bestuur en ontleding van grootdata verteenwoordig.

Binne die geesteswetenskappe is daar 'n ander formaat wat bygereken behoort te word by dataformate: analoog (hardekopie). Alhoewel daar groot hoeveelhede inligting in 'n digitale formaat beskikbaar is, word die navorser binne die geesteswetenskappe gereeld gekonfronteer met dokumente wat slegs in hardekopie beskikbaar is (byvoorbeeld ouer Afrikaanse literêre werke of argiefmateriaal). Verwerking sluit dan vir die navorser in die geesteswetenskappe nie alleen in die omskakeling van ongestruktureerde na gestruktureerde data nie, maar soms ook die digitalisering van bronmateriaal. Dít verg bykomende tegnologiese vaardigheid, programmatuur en apparatuur, asook 'n verdere tydinset. Die datastel wat soms in hierdie boek gebruik word rakende die Afrikaanse literêre sisteem vanaf 1900 tot 1978 is juis deur so 'n lang verwerkingsproses wat begin het by die digitalisering van die brondokumente (Senekal en Van Aswegen (1980, 1981) en Senekal en Engelbrecht (1984)), hierna omgeskakel is vanaf 'n ongestruktureerde formaat (PDF) na 'n gestruktureerde formaat (Microsoft Excel), en ook skoongemaak is om te verseker dat data konsekwent ingevoer is. Selfs al is bronmateriaal reeds in 'n digitale formaat, gebeur dit soms by ouer dokumente dat karaktererkenning nie toegepas is nie (byvoorbeeld koerantuitknipsels), wat dan beteken dat die navorser eers karaktererkenning sal moet toepas voor hy enige inligting sal kan onttrek. Digitalisering behels nie alleen die skandering van dokumente nie, maar ook verheldering, die toepassing van karaktererkenning en dergelike. In die geval waar karaktererkenning nie reeds toegepas is nie sal die navorser hierdie stap, wat by die digitaliseringsproses tuishoort, moet toepas. Alhoewel digitalisering buite die bestek van hierdie boek val, sal die navorser ook baat vind daarby om hiermee bekend te word en die nodige infrastruktuur aan te skaf, soos bespreek in Senekal (2011).

1.4 Gevolgtrekking

Grootdata kan in navolging van Russom (2011:6) opgesom word soos in Figuur 2.



Figuur 2. Aspekte van grootdata

Grootdata is data wat groot hoeveelhede inligting beslaan (soos gemeet in die aantal grepe of rekords) of 'n hele datastel verteenwoordig, deurlopend gegenereer en gereeld in reële tyd ontleed word, en 'n verskeidenheid formate aanneem. Vir die geesteswetenskappe is veral die grootte en verskeidenheid van data van belang, en later in hierdie boek sal oplossings bespreek word wat betrekking het op hoe hierdie uitdagings die hoof gebied kan word. Eerstens is dit egter belangrik om ondersoek in te stel na die implikasies wat grootdata vir die wetenskap inhou, soos in die volgende hoofstuk bespreek word.

Hoofstuk 2

Die implikasies van grootdata vir die wetenskap

Boyd en Crawford (2012:663) skryf dat die diskoers rondom grootdata beide utopiese as distopiese retoriek ontketen, soos Loader en Dutton (2012:610) skryf ook die geval is met die internet self. Mayer-Schönberger en Cukier (2013:19) (wat vanuit 'n utopiese invalshoek skryf) stel voor dat 'n aantal belangrike kopskuiwe gemaak moet word in die oorgang van tradisionele wetenskaplike metodes na die ontleding van grootdata, soos in hierdie afdeling bespreek. Heelwat van hierdie aspekte is aanvegbaar en reeds breedvoerig gekritiseer, en kritiek op die grootdatakonsep en -bewerings word ook onder die loep geneem.

2.1 Die einde van steekproefneming

Die eerste groot metodologiese skuif wat grootdata veronderstel is dat steekproewe onnodig word: grootdata kyk na 'n datastel as geheel, nie 'n ewekansige of verteenwoordigende monster van 'n geheel nie (Mayer-Schönberger & Cukier 2013:20-31; Davenport 2014:94; Jockers 2013:7). Steekproefneming is volgens Mayer-Schönberger en Cukier juis 'n kompromis wat in die verlede aangegaan is omdat datastelle nie in geheel versamel of ontleed kon word nie; 'n hele bevolking van 'n paar miljoen mense kon immers nie ondervra word nie (behalwe in die geval van 'n sensusopname). Die nadele van steekproefneming is aldus die outeurs tweeledig: indien monsters nie korrek geselekteer word nie, kan groot foute in 'n studie insluip, en ook verdwyn die resolusie, wat beteken dat individuele gevalle verlore gaan in die ontleding.

Steekproefneming kom met risiko's wat die geldigheid van 'n navorsingprojek bedreig. Watts (2011:113) skryf dat mense in die alledaagse wêreld meer aandag skenk aan interessante gebeurtenisse as oninteressante gebeurtenisse, wat hy 'n steekproefvooroordeel (*sample bias*) noem. 'n Mens sal byvoorbeeld let op al die kere wat jy ou skoolvriende op onwaarskynlike plekke raakloop, of treine verpas, maar nie let op al die kere wat dit *nie* gebeur het nie. Dít is natuurlik 'n terloopse waarneming en geen wetenskaplike studie nie, maar akademici is nie verhewe bo sulke steekproefvooroordele nie. Na die Universiteit van die Vrystaat se Reitz-video insident is die video deur sommige

mense in verband gebring met die Waterkloof Vier en die Skierlik-skietvoorval, en saam gesien as voorbeelde van rassisties-gemotiveerde aanvalle van wit mense op swart mense. Dié 'steekproef' laat egter buite rekening dat daar elke dag miljoene interaksies tussen wit en swart in Suid-Afrika is wat geen probleme veroorsaak nie, maar dit is natuurlik nie vir die media interessant om te let op goeie interaksies nie. Die risiko van só 'n fout kan verminder word deur 'n groter monster te neem (vier individue is 'n baie klein monster en onvanpas vir die gevolgtrekking wat gemaak is), deur die monster ewekansig te selekteer (ewekansig geselekteerde monsters lewer gewoonlik meer geldige resultate as verteenwoordigende monsters (Mayer-Schönberger & Cukier 2013:22)), en deur nie terloopse voorbeelde te selekteer nie, maar sistematies te werk te gaan. Foute kan uiteraard steeds in enige steekproefneming insluip, wat die uiteindelike geldigheid van die studie ondermyn.

Die siening van die letterkunde as 'n aantal verteenwoordigende tekste (die kanon) sluit ook hierby aan omdat dit 'n steekproef veronderstel. Dié benaderingswyse word veral gekritiseer deur Jockers (2013) en Moretti (2005). Moretti (2005:4) let daarop dat 'n kanon van 200 romans bestudeer word as verteenwoordigend van die Britse literatuur van die 19de eeu, maar dat dit minder as 1% van die gepubliseerde werke verteenwoordig (daar is meer as 20 000 romans in dié eeu in Brittanje gepubliseer). Om die 'kenmerke' van die negentiende-eeuse Britse roman uit so 'n klein monster te veralgemeen kom uiteraard met probleme. Die probleem word boonop vererger deurdat die monster nie ewekansig óf verteenwoordigend geselekteer is nie; gewoonlik vind seleksie plaas op grond van literatuurhistorici se waardeoordele, wat om die beurt deur hul literatuuropvattinge beïnvloed word. Moretti en Jockers bepleit 'n inklusiewe literatuurgeskiedenis om hierdie steekproefvooroordeel te korrigeer, maar só 'n inklusiewe siening van die literêre sisteem maak dit onmoontlik om sonder rekenaarprogrammatuur met die letterkunde om te gaan.

Verder verdwyn die resolusie in gewone steekproefnemings. Wanneer 'n ewekansige monster van 'n bevolking geneem en veralgemeen is na 'n bevolking, kan die individu se posisie in die datastel nie agterna gevind word nie. So het die jongste sensus byvoorbeeld aangedui dat 29,8% van Bloemfonteiners wit is, en 42,5% van die hele bevolking Afrikaans is teenoor 7,5% Engels. Dié statistiek beskryf nie een van die outeurs van hierdie boek se onmiddellike sosiale kontekste nie (dit is byvoorbeeld nie die geval dat 7,5% van ons vriende Engels is nie), en ook kan die statistiek nie gebruik word om die individu se posisie te bepaal nie. Steekproefneming is van 'n lae resolusie: indien 'n navorser wil 'afboor' na individuele gevalle, verdwyn die inligting. 'n Mens kan in hierdie opsig aan 'n digitale kaart soos Google Maps dink: op elke hoogte gee die kaart nie *al* die inligting wat beskikbaar is oor 'n gebied nie, maar slegs dit wat

relevant is. Indien 'n mens egter afboor en die kaart vergroot, word details ingevul wat nie op 'n groter afstand weergegee is nie. 'n Steekproefneming is dus soos 'n kaart in hardekopie; daar kan nie tussen vlakke van gegewens beweeg word nie, in teenstelling met die grootdatabenadering wat soos 'n interaktiewe digitale kaart funksioneer.

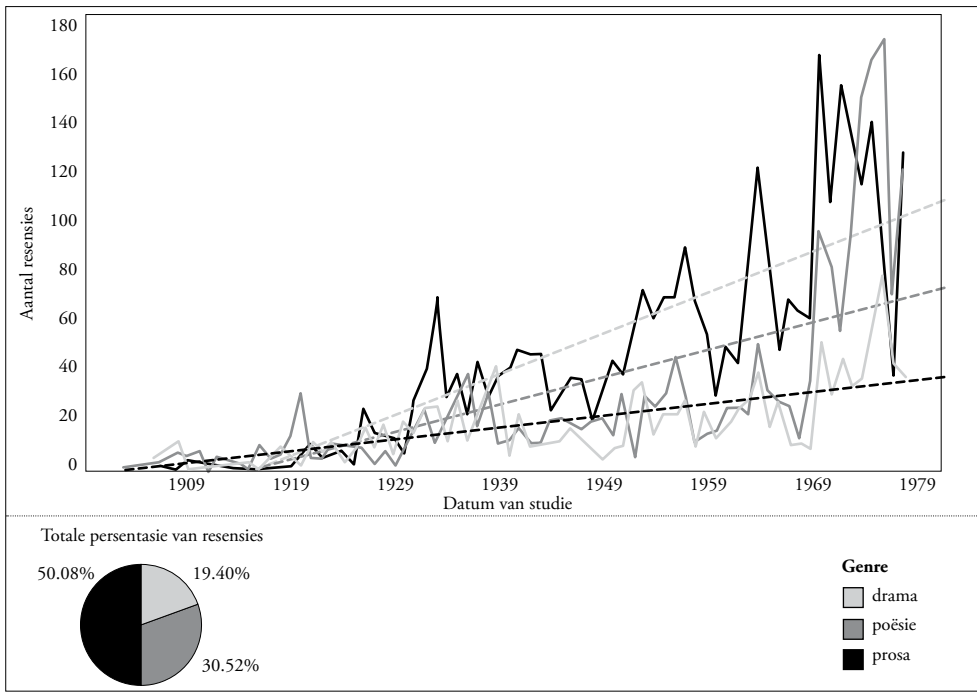
Jockers (2013:7) skryf ook dat uitsonderings nie behoorlik in steekproefneming na vore kom nie; in steekproefneming verdwyn gevalle rondom die gemiddeld. Dit is byvoorbeeld wel so, indien 'n mens na historiese verskiesingsuitslae kyk, dat die meerderheid wit mense die apartheidsregering gesteun het, maar hierdie gemiddeld verreken nie die uitsonderings soos Bram Fischer, Beyers Naudé, Breyten Breytenbach en al die ander ondersteuners van opposisiepartye nie (dieselfde geld vir swart mense, van wie sommiges ook die apartheidsregering gesteun het). Hierteenoor kyk grootdata na die hele datastel op so 'n wyse dat die uitsonderings steeds sigbaar is, omdat 'n grootdatabenadering 'n hoë resolusie het.

Resolusie is dus belangrik in hierdie verband. Soos Mayer-Schönberger en Cukier (2013:160-161) se bogenoemde stelling dat grootdata die individu kan bevry van sy groepsidentiteit, kan grootdata juis van veel waarde wees in Suid-Afrika, waar rassevooroordele vanuit verskeie oorde mense steeds in terme van groepsidentiteite sien. Groepsidentiteite is juis die gevolg van 'n swak resolusie en tradisionele statistiek. Hermann (2013) teken sterk protes aan teen die huidige regering se statistiese benadering tot regstellende aksie wat individuele gevalle ignoreer. Onses insiens is die duidelikste voorbeeld van 'n laeresolusieprobleem wat uit Hermann se boek na vore kom, dié van Christo February, wat as anti-apartheidsaktivis nou *benadeel* word deur regstellende aksie (2013:72-75). Die huidige regering maak juis gebruik van tradisionele statistiek en groepsidentiteite om te bepaal wie werk moet kry, en, soos Hermann, bepleit die grootdatabenadering ook dat die individu wat in die statistiek vervat is, belangrik is.

Grootdatabenaderings se vermoë om af te boor na individuele gevalle is veral belangrik wanneer tekste ontleed word, soos Jockers (2013:23; 2014:vii) aanvoer. Hy (2014:vii) skryf dat rekenaarmatige berekenings toegang bied tot inligting in tekste wat 'n mens eenvoudig nie kan versamel met tradisionele kwalitatiewe metodes van noukeurige lees en menslike sintese nie; die beloning lê volgens hom daarin dat rekenaarmatige ontledings toegang tot inligting bied op beide die makro- en mikroskaal. Elders (2013:89 e.v.) illustreer hy die waarde van 'n grootdatabenadering wanneer hy woorde identifiseer wat veral saamhang met sekere genres, byvoorbeeld met betrekking tot die Bildungsroman. Na die identifisering van sulke woorde in 'n makroanalise, boor hy af na die individuele woorde en die konteks waarin hul voorkom, en dui aan hoe dié woorde temas van genres vergestalt. Een woord wat gereeld in die Bildungsroman voorkom is 'like', en by nadere ondersoek vind Jockers dat dit saamhang met die Bildungsroman se

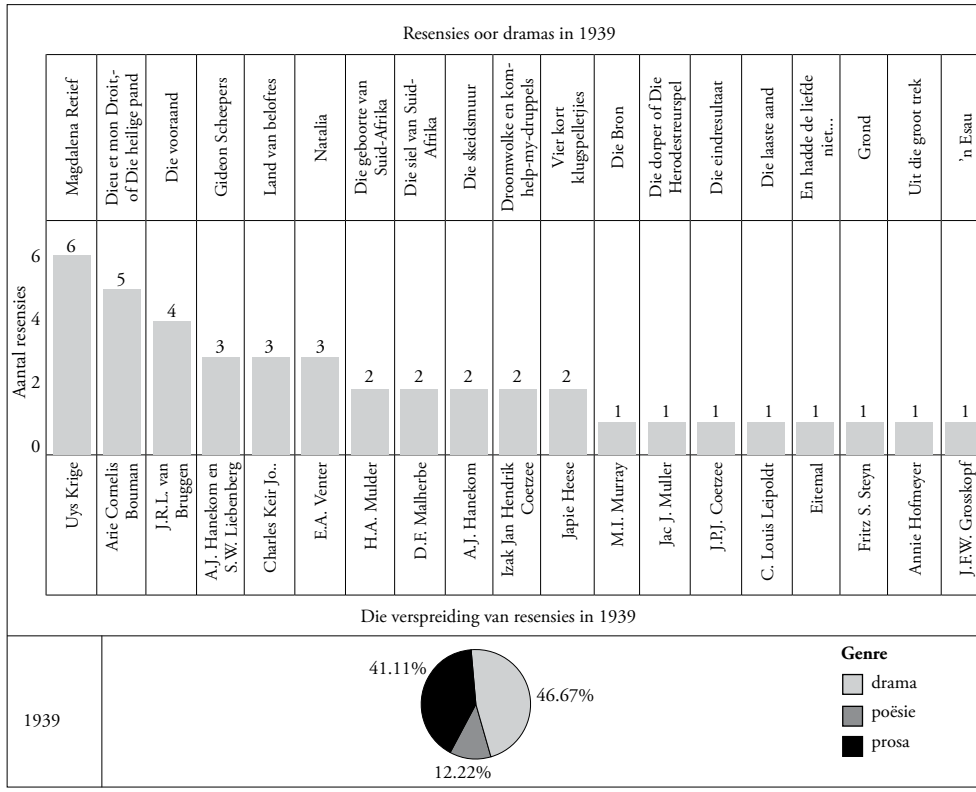
ontdekking van die volwasse wêreld wat dan vergelyk word met die bekende wêreld van die kind. Die voorkoms van dié woord dui dan spesifiek op die Bildungsroman se posisie tussen die wêreld van die kind en die volwassene. Só faset van 'n genre kan natuurlik nie uitgelig word wanneer slegs enkele, 'verteenwoordigende' tekste bestudeer word nie, en die grootdatabenadering bemagtig die navorser om meer indringend na sy studie-objek te kyk. Belangriker nog is die feit dat die navorser wel 'n breë oorsig oor die korpus kan onderneem, maar dan steeds afboor na die individuele gevalle, aangesien die individuele datapunt se posisie steeds opgespoor kan word.

Neem byvoorbeeld die Afrikaanse literêre sisteem vanaf 1900 tot 1978, met data verkry vanuit Senekal en Van Aswegen (1980, 1981) en Senekal en Engelbrecht (1984). Die datastel wat hieruit saamgestel is, bestaan uit meer as 110 000 datapunte, wat 'n omvattende oorsig bied oor wie wat in die Afrikaanse letterkunde gepubliseer het (drama, poësie en prosa), wie wat oor hierdie werke gepubliseer het (resensies, studies en literatuurgeskiedenisse), sowel as waar hierdie studies en resensies verskyn het. Op 'n makrovlak kan daar byvoorbeeld aangedui word watter genres die meeste aandag van kritici ontvang het oor die hele tydperk, soos gesien in Figuur 3 (die verkenning is met behulp van Tableau onderneem).



Figuur 3. Die verspreiding van studies in die Afrikaanse letterkunde

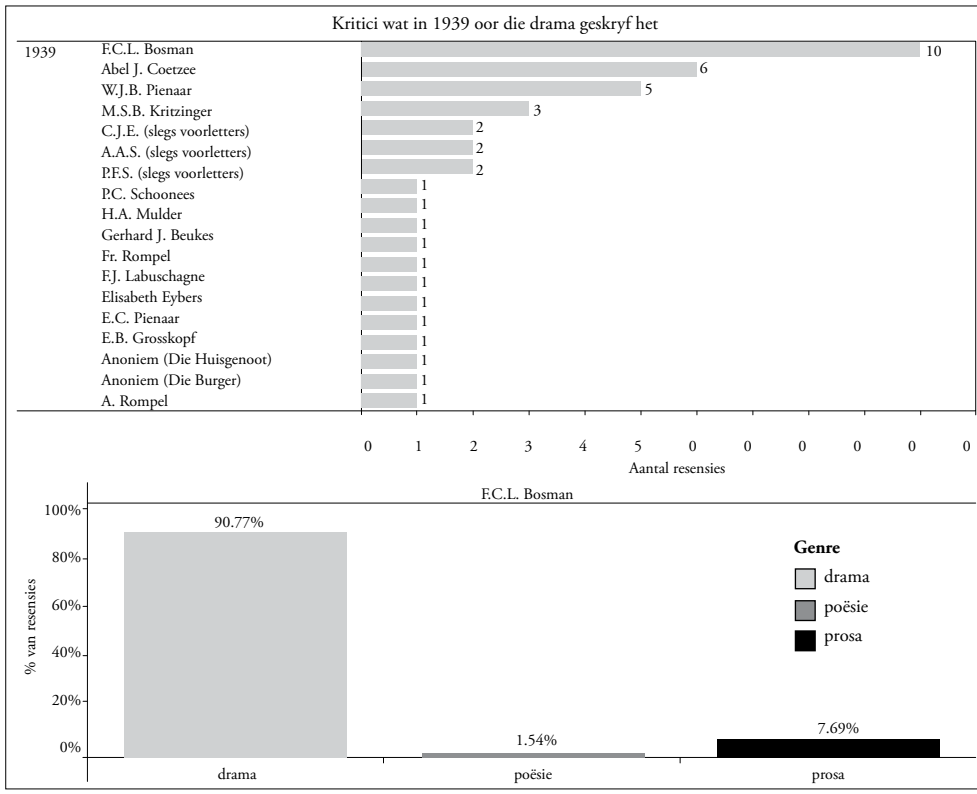
Die meerderheid resensies (50.08%) handel hiervolgens oor prosa. Dit is logies dat die aandag wat kritici aan die onderskeie genres bestee van tyd tot tyd sal verskil, en daarom is die verspreiding van resensies oor dié drie genres aangedui vir die tydperk, met gemiddelde waardes ook aangedui deur stippellyne. Hier kan duidelik gesien word dat daar sedert 1919 gemiddeld die minste oor die drama gepubliseer is, maar dat daar uitsonderings in sommige jare bestaan. Soos in Figuur 4 gedemonstreer handel die meerderheid publikasies in 1939 oor dramas, wat daartoe gelei het dat drama dié jaar se literêre diskoers oorheers het.



Figuur 4. Dramas wat in 1939 bestudeer is

Dit was dus veral Uys Krige se *Magdalena Retief* (1938) en Arie Cornelis Bouman se *Die heilige pand* (1939) wat in dié jaar aandag van 'n groot aantal kritici ontvang het, gevolg deur *Die vooraand* (1939) van J.R.L. van Bruggen (let wel dat datums hier dui op

wanneer daar oor die betrokke werke geskryf is, nie wanneer dié dramas gepubliseer is nie). Figuur 5 dui aan wie gedurende 1939 die meeste oor drama geskryf het.



Figuur 5. Letterkundiges wat in 1939 oor die drama gepubliseer het

F.C.L. Bosman was duidelik die letterkundige wat in dié jaar die grootste aantal dramas geresenseer het, gevolg deur Abel J. Coetzee – hier kan ook gesien word dat die meeste van Bosman se resensies oor die algemeen oor die drama handel. In die voorafgaande voorbeeld is daar dus gewerk vanaf 'n breë makrovlakoorsig wat die hele datastel in ag geneem het deur na die individu se posisie in die datastel te kyk, en die enkele datapunt se posisie kan dus in die breër geheel nagespeur word. Dit is 'n beduidende voordeel van grootdatabenaderings; wanneer 'n steekproef na 'n geheel veralgemeen word kan individuele data nie soos hier ondersoek word nie.

Steekproewe gaan volgens Mayer-Schönberger en Cukier (2013:30) altyd gepaard met die risiko dat mense anders sal reageer in 'n navorsingsopset as in hul alledaagse lewe, iets wat soms na verwys word as die sogenaamde Hawthorne-effek.

Henry A. Landsberger het in 1950 vroeëre eksperimente by die Hawthorne fabriek naby Chicago ontleed, en voorgestel dat die deelnemers aan die studie anders reageer omdat hulle besef hulle word bestudeer.[16] Grootdata kan só 'n gevolg voorkom deur byvoorbeeld selfoondata te benut om mense se kommunikasiepatrone na te speur soos dit in hulle werklike, alledaagse lewens manifesteer. Barabási het byvoorbeeld selfoondata bekom vir 100 000 gebruikers in 'n onbekende Europese land, en was in staat daartoe om hul bewegings oor ses maande te monitor. Die vraag is egter of die gemiddelde akademiese navorser toegang tot sulke data kan verkry, en natuurlik word sulke voordele van grootdata beïnvloed deur kwessies van privaatheid.

Alhoewel Davenport (2014:94) saamstem dat steekproefneming minder belangrik word in 'n era van grootdata, stel hy die klemverskuiwing versigtiger as Mayer-Schönberger en Cukier. Hy voer aan dat dit nie geheel en al oorbodig word nie (dit bly byvoorbeeld 'n probleem om die hele bevolking van 'n land oor 'n spesifieke kwessie te raadpleeg), maar steekproefneming word wel minder belangrik. Benewens 'n grootdatabenadering benut eBay steeds bykomende steekproefneming om hul marknavorsing te doen (Davenport 2014:164). Dit is ook betwyfelbaar of alle vrae in die wetenskap met behulp van omvattende datastelle beantwoord kan word. 'n Mens sou kon sê dat Mayer-Schönberger en Cukier se stelling dat grootdata die einde van steekproefneming meebring ietwat oordrewe is, maar dat die klem al hoe meer verskuif na die ontleding van datastelle in die geheel.

2.2 Die einde van presiese datastelle

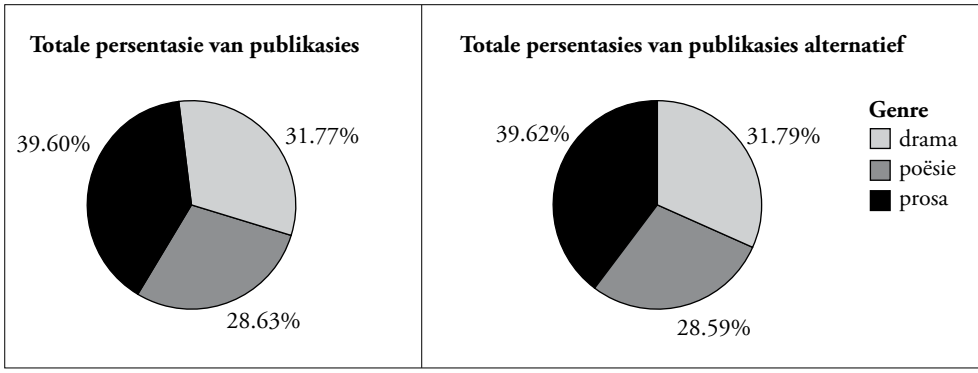
Verder bring grootdata ook mee dat die akkuraatheid van datastelle vervang word met omvattendheid. Mayer-Schönberger en Cukier (2013:32-49) beweer dat akkuraatheid noodsaaklik was in 'n tyd toe daar minder data beskikbaar was en die gereedskap nie beskikbaar was om groot hoeveelhede data mee te ontleed nie. Volgens die outeurs is foute in grootdatastelle onvermybaar en bedreig dit nie die geldigheid van die studie nie.[17] Agrawal et al. (2011:6) wys daarop dat die grootte van datastelle hierdie foute uitskakel, deels omdat gapings in die datastelle deur oorbodigheid gevul word.

Hierdie stelling is waarskynlik kontroversieel, en sal daarom deur 'n voorbeeld geïllustreer word. Met behulp van Senekal en Van Aswegen (1980, 1981) en Senekal en Engelbrecht (1984) kan 'n lys van die 1 682 Afrikaanse literêre publikasies wat tussen

16 Sien byvoorbeeld Landsberger (1958).

17 Sien ook Loukides (2010:5).

1900 en 1978 uitgegee is saamgestel word. Dié werke is soos volg oor genres versprei: 39,6% prosa, 31,77% drama, en 28,63% poësie. Senekal en Van Aswegen (1980, 1981) ressorteer N.P. van Wyk Louw se *Die dieper reg* (1938) onder beide poësie en drama, aangesien dié werk 'n versdrama is. Indien dié werk egter slegs as 'n drama geag is, sou die verspreiding van werke tussen genres anders daar uitsien, soos in Figuur 6 aangedui.



Figuur 6. Die verspreiding van werke in die Afrikaanse letterkunde tussen 1900 en 1978

Die individuele werk het dus 'n invloed op die statistiek, alhoewel 'n baie klein een: die poësie maak nou 0.04% minder van die datastel uit, terwyl die drama en die prosa elk 0.02% meer uitmaak. Dit is 'n geringe verskil, maar onthou dat hierdie 'n relatiewe klein datastel is soos gemeet ten opsigte van die aantal rekords wat betrokke is (dit is grootdata in die sin daarvan dat dit 'n omvattende datastel is). Wanneer daar miljoene, eerder as 1 682, rekords by 'n studie betrokke is, word die invloed van 'n enkele geval weglaatbaar klein. 'n Tik- of spelfout, en foutiewe of ontbrekende inligting beïnvloed nie die eindresultaat nie solank dit beperk is tot enkele gevalle.

Barabási (2011:14) skryf dat groot datastelle wel soms onvolledig mag wees, maar dat netwerkontledings steun op 'skerp toerusting' – met ander woorde 'n baie presiese navorsingsmetode – om hierdie datastelle te ontleed. Dit bring mee dat die metode self ook vergoed vir foute in die data. Netwerkontledings is veral bekend daarvoor dat dit vergewend is teenoor geringe foute in die onderliggende data: in 'n netwerk met tienduisende of miljoene entiteite kan 'n enkele entiteit nie die gemiddelde pad tussen alle entiteite beduidend beïnvloed nie. Neem byvoorbeeld die internasionale filmakteurnetwerk, wat in 'n verskeidenheid studies bestudeer is. Die gemiddelde pad tussen akteurs behels die aantal skakels wat enige akteur met enige ander akteur verbind, soos in Tabel 4 weergegee (aantal nodusse of akteurs n en gemiddelde pad l is aangedui).

Tabel 4. Die gemiddelde pad in akteurnetwerke

Netwerk	**n**	**l**	**Studie**
Akteurs	225 226	3,65	Watts en Strogatz (1998)
Akteurs	212 250	4,54	Barabási en Albert (1999)
Akteurs	449 913	3,48	Amaral et al. (2000)
Akteurs	392 340	3,6	Guillaume en Latapy (2006)
Akteurs	127 823	6,8	Latapy, Magnien, en Del Vecchio (2008)
Akteurs	392 340	3,6	Guillaume en Latapy (2004)

Al hierdie studies maak gebruik van die Internet Movie Database (www.imdb.com), wat wel die mees omvattende datastel rakende die internasionale filmindustrie is, maar nogtans onvolledig is – veral ten opsigte van Afrikaanse films en films wat nie groot internasionale suksesse was nie. *Arende* (1994) het byvoorbeeld volgens www.imdb.com slegs 3 akteurs, terwyl die film 27 akteurs in die krediete lys. Een van die outeurs (Senekal) het bereken dat wanneer die gemiddelde pad vir die akteurnetwerk binne die Afrikaanse filmindustrie sedert 1994 uitgewerk word met akkurate data wat vanuit die films self verkry is, die gemiddelde pad 2,33 vir die 1 715 akteurs wat by dié industrie betrokke is. Dit kan verwag word dat die Afrikaanse filmindustrie, wat natuurlik baie klein is in vergelyking met die internasionale filmindustrie, 'n korter pad as die algehele gemiddeld sal hê, veral omdat daar aansienlik minder films en akteurs by dié industrie betrokke is en akteurs dus meer gereeld saamwerk. Bogenoemde studies het 'n kort gemiddelde pad uit groot datastelle geïdentifiseer, en dit is inderdaad merkwaardig dat daar 449 913 akteurs in Amaral et al. (2000) se studie is wat gemiddeld slegs 3,48 skakels van mekaar verwyderd is, maar die feit dat 'n gemiddelde kortpad hierdie netwerk kenmerk word nie beïnvloed as 'n mens met akkurater data werk nie (3,48 met inagname van 449 913 akteurs is kort, soos 2,33 vir die 1 715 akteurs in die Afrikaanse filmakteurnetwerk ook kort is). Dit is wat Barabási bedoel met 'skerp' toerusting: teorie en wiskundige formules wat só akkuraat is dat dit vergewensgesind is teenoor onvolledige data.

2.3 Die einde van kousaliteit

Verder val die klem in grootdata op korrelasie eerder as kousaliteit[18] (Mayer-Schönberger & Cukier 2013:50-72). 'n Voorbeeld is Amazon se voorstelle van wat om ook te koop as 'n mens reeds een ding gekoop het: dié voorstelle is gegrond op gesofistikeerde algoritmes wat ondersoek instel na korrelasies in verbruikers se kooppatrone. *Hoekom* iemand wat in produk X belangstel ook in Y sal belangstel, is irrelevant; die oorweging is dat die *wat* belangrik is (Mayer-Schönberger & Cukier 2013:52). Dié algoritmes het die aanvanklike groep kritici wat Amazon aangestel het om resensies te skryf uiteindelik vervang, omdat die kritici nie korrek kon voorspel waarin mense ook sou belangstel nie. Volgens Mayer-Schönberger en Cukier (2013:61) is die voordeel van grootdata in hierdie opsig dat dit vooroordele teenwerk, aangesien die gevaar nie bestaan dat 'n navorser met vooropgestelde idees na 'n onderwerp gaan nie – die data 'spreek vanself' (kritiek op hierdie siening word later bespreek).

Kousaliteit is reeds 'n problematiese konsep in die wetenskap. Watts (2011:199) herinner dat dit onmoontlik is om kousaliteit wetenskaplik te bepaal sonder eksperimente, wat 'n vraagteken plaas oor enige bewerings van oorsaaklikheid in byvoorbeeld die geskiedenis, waar eksperimente onmoontlik is. Net omdat B volg op A beteken geensins dat daar 'n kousale verband is tussen A en B nie (Watts 2011:118). Watts (2011:116-117) herinner aan die siening dat skoolkinders wat by skietvoorvalle betrokke was, vervreem is van hul portuurgroep en/of familie, en blootgestel is aan gewelddadige videospeletjies en televisieprogramme, en dat daar 'n kousale verband tussen hierdie faktore en hul latere optrede beweer word. Dié beweerde kousale verband laat egter buite rekening dat daar letterlik miljoene tieners is wat óók vervreem is van hul omgewing en óók blootgestel is aan geweld deur die media, maar wat nooit enigeen skiet nie. Die kousale verband is dus vals, maar omdat die een die ander volg, kom dit voor of daar wel 'n kousale verband bestaan. In 'n Suid-Afrikaanse konteks kan die voorval van die Reitz-video weer genoem word. Net omdat mense in ouerhuise grootgeword het wat hul oorsprong onder apartheid gehad het, beteken nie noodwendig dat mense rassistiese houdings sal huldig nie, en verseker nie dat hulle noodwendig betrokke sal wees in rassisties-gemotiveerde aanvalle nie. Daar is egter sommige akademici wat beweer dat dit wel die geval is; dat apartheid die enigste oorsaak van sogenaamde rassistiese voorvalle is, en dit word ook soms in deterministiese terme gestel dat dit jongmense se opvoeding in ouerhuise is wat lei tot hierdie voorvalle. Daar is egter honderde duisende wit jongmense

18 George, Haas, en Pentland (2014:323) skryf egter dat kousaliteit steeds bestudeer moet word nadat korrelasies van fenomene geïdentifiseer is.

wat *nie* betrokke is of was by rassistiese voorvalle nie, maar wat ook uit huise kom wat hul oorsprong in die apartheidsbestel gehad het, wat beteken dat die deterministiese kousaliteit wat voorgestel word nie deur die werklikheid onderskraag word nie.

Hiervolgens is kousaliteit ook 'n probleem in die letterkunde. Wanneer Jockers (2013:108 e.v.) skryf oor die feit dat die gebruik van die bepaalde lidwoord in Amerikaanse en Britse tekste gekorreleer is, met ander woorde wanneer die gebruiksfrekwensie van die bepaalde lidwoord in Amerikaanse tekste toeneem, neem dit ook toe in Britse tekste, spekuleer hy oor moontlike redes hiervoor, maar laat vaar uiteindelik sy soektog na kousaliteit en aanvaar bloot die korrelasie (wat hy spesifiek aandui nie bloot toevallig is nie). Daar bestaan derhalwe 'n definitiewe, aantoonbare en bewysbare verhouding tussen die Britse en Amerikaanse literêre sisteme bloot op grond van die korrelasie van die gebruiksfrekwensie van die bepaalde lidwoord, maar Jockers – wat homself deurgaans by die wetenskaplike metode hou – erken uiteindelik dat hy nog geen kousale verband kon identifiseer nie.

In dieselfde opsig is dit moeilik om te bepaal waarom 'n teks werklik gesien word as 'n meesterwerk. Omdat literatuur nie in isolasie funksioneer nie, speel literatuuropvattinge, ideologie, leserinterpretasies en dies meer ook 'n rol. Watts se eksperimente met musiek (Salganik & Watts 2008; Watts & Hasker 2006) het aangedui dat die gewildheid van 'n liedjie bo alle twyfel beïnvloed word deur sosiale faktore: wanneer mense weet dat 'n liedjie gewild is, neem hulle aan dat dit goed moet wees. 'n Mens sou ook in die letterkunde kon postuleer dat wanneer 'n teks of skrywer gewild in akademiese kringe is (met ander woorde hoog aangeskrewe staan), daar 'n konsensus ontwikkel wat aandui dat dit 'n 'goeie' teks of skrywer is, wat om die beurt die siening van die teks beïnvloed. Watts skryf egter dat dit nie bloot sulke ekstrinsieke faktore is wat waardeoordele beïnvloed nie, maar ook intrinsieke eienskappe van die teks: 'goeie' liedjies het in sy eksperiment deurgaans goed gevaar, terwyl 'swak' liedjies deurgaans swak gevaar het (Watts 2011:77). Uit Watts se studies kan juis gesien word hoe daar 'n wisselwerking tussen die intrinsieke en ekstrinsieke ontstaan, en dit is onmoontlik om te bepaal wat uiteindelik veroorsaak dat 'n liedjie gewild is, juis omdat 'n kombinasie van faktore vir gewildheid verantwoordelik is.[19]

Die primêre probleem met kousaliteit is dat die menslike beweegruimte 'n komplekse sisteem is waar die geheel meer as die somtotaal van die onderdele is. Page (2011:217) stel dié beginsel in wiskundige terme: $f(x+y) \geq f(x) + f(y)$. Dit beteken dat 'n funksie f van die sisteem nie alleen die funksie van element x plus die funksie van element

19 Watts (2011) skryf ook in soortgelyke terme oor Harry Potter en die Mona Lisa.

y is nie, maar heelwat meer. Christakis en Fowler (2010:26) let daarop dat die smaak van 'n koek meer is as die somtotaal van die smake van sy bestanddele; 'n koek proe immers nie soos eiers plus meel nie.[20] Wanneer elemente in 'n sisteem in 'n interafhanklike verhouding tot ander elemente staan, word dit uiters moeilik om oorsaak en gevolg te bepaal (Holland 2006:2), en aangesien die menslike beweegruimte – juis die studie-objek binne die geesteswetenskappe – binne 'n interafhanklike netwerk van komplekse sisteme bestaan (DeLaurentis 2007:363), word hierdie probleem op die spits gedryf binne die geesteswetenskappe. 'n Rassisties-gemotiveerde insident is dus nie eenvoudig die gevolg van 'n persoon se opvoeding plus die hedendaagse realiteit van rasse-integrasie nie, maar ook nie eens 'n meer komplekse formulering soos opvoeding plus regstellende aksie plus misdaad plus integrasie nie. Só 'n insident is die gevolg van veel meer faktore (onder andere ekonomies, sosiaal en sielkundig), en belangriker nog: dit is die gevolg van die *kombinasie* van faktore. Mense tree nie in isolasie op nie, maar is ingebed binne 'n komplekse web van interaksies wat die media, politici, families, vriende, kollegas, kennisse, hul eie psiges en die chemiese interaksies in hul breine, gemeenskaplike norme en ideologieë, asook ekonomiese, politieke en maatskaplike sisteme insluit. Om in só 'n komplekse omgewing soos die menslike beweegruimte 'n enkele kousale verband voor te hou, is wetenskaplik ongeldig, en om die presiese interafhanklike wisselwerking van alle faktore na te speur is onmoontlik.

Anderson (2008) en ander datawetenskaplikes se oplossing vir hierdie probleem is om kousaliteit te ignoreer en eerder op korrelasie te fokus. Dié klemverskuiwing beteken nie dat kousaliteit nie bestaan nie, maar eerder dat die probleme rondom die wetenskaplike bewys van kousaliteit in veral die menslike beweegruimte kousaliteit se waarde binne geesteswetenskaplike navorsing bevraagteken. Korrelasie is meer geredelik bewysbaar binne die geesteswetenskappe.

2.4 Die einde van teorie

Een van die bewerings wat gereeld aangehaal word in die diskoers rondom grootdata is Anderson (2008) se veronderstelling dat ons by die einde van teorie gekom het in die wetenskap: grootdata kan in hierdie siening bevindinge uit die data genereer sonder die opstelling van 'n hipotese of om binne 'n teoretiese raamwerk te werk. Anderson sê dat daar weggedoen moet word met elke teorie rakende menslike gedrag, van die linguistiek tot die sosiologie. Volgens hom kan daar ook van die sielkunde ontslae geraak

20 Sien ook Kilcullen (2010:195) en Nicolis (1995:1-2).

word aangesien dit onbelangrik is *hoekom* mense doen wat hulle doen – wat tel is *wat* hulle doen. Dít kan volgens hom uit die data gesien word, wat 'vanself spreek.' Hierdie siening bied glo die voordeel dat die navorser nie die risiko loop om gelei te word deur sy teoretiese raamwerk nie.

Eerstens moet 'n mens gelyk gee aan Anderson en ander datawetenskaplikes en erken dat baie van wat aangebied word as 'wetenskap' (veral in die geesteswetenskappe) geensins wetenskaplik is nie, veral wanneer akademici gelei word deur 'n bepaalde ideologie of deur teorieë wat nie op empiriese navorsing gegrond is nie. Die diskoers rondom Jamie Uys se films kan as voorbeeld dien: byna geen studie is nog oor Uys se films gepubliseer wat nie aanvoer dat dié films rassisties is nie (Senekal en J.-A. Stemmet 2014:1). 'n Keurder het voorgestel dat Senekal en Stemmet ook hierdie onderwerp aanspreek, alhoewel Uys se beweerde onderliggende ideologie geensins relevant is vir 'n ontleding van sy samewerkingsnetwerk nie ('n opmerking hieroor is noodgedwonge in die finale weergawe van die artikel ingesluit). Die studie van outeursbedoeling is reeds lank terug opsy geskuif in die literatuurstudie omdat dit onwetenskaplik is (Senekal 1987:52). Om aan te voer dat 'n mens Uys se houding jeens anderskleuriges uit sy films kan agterhaal is ongegrond. Ook is die kyker (of leser van die filmdraaiboek as teks) se interpretasie geensins neutraal nie, en moet dit verreken word dat daar 'n verskil is tussen hoe 'n teks of film geïnterpreteer word en wat 'n wetenskaplike, verifieerbare feit veronderstel. Senekal (1987:19) skryf:

> Die interpretasie van tekste, uitsprake waarin geprobeer word om die 'betekenis' van 'n literêre teks weer te gee, is nie vatbaar vir toetsing nie. Hulle bestaan ook nie as afsonderlike entiteite nie. Die eienskappe wat aan 'n teks toegeken word, asook die verbande wat die interpretasies ten grondslag lê, word gedoen op grond van die persoonlike belewing van die interpreteerder. Resultate wat op hierdie manier verkry word, is uitgesluit van vrywel enige vorm van verdere diskussie.

Volgens Senekal is só 'n interpretasie-gebaseerde studie uiters problematies: "Die ondersoeker verkeer onder die indruk dat hy relevante feite bestudeer. Wat nou gebeur, is dat wetenskaplike status toegeken word aan lesersuitsprake". Die diskoers rondom Jamie Uys se films dryf juis dié geneigdheid om van lesersuitsprake wetenskaplike feite te maak op die spits: dit is al 'n aanvaarde 'feit' in filmstudies dat Uys se films rassisties is. Oor die onwetenskaplikheid van sulke lesersuitsprake is Senekal (1987:27) ondubbelsinnig: "Met deelname aan die literatuursisteem, bv. op die vlak van die essayistiek, die interpretasie, die kritiek, kan niemand probleme hê nie. Solank dit geen wetenskaplike pretensies

het nie". Die probleem is egter dat sulke lesersuitsprake wel as wetenskap aangebied word, met die gevolg dat die geesteswetenskappe se reputasie as wetenskaplike dissipline ondermyn word. Voeg hierby steekproefvooroordele en onvanpaste veralgemenings soos hierbo in die konteks van die Reitz-video bespreek, en 'n mens ontwikkel begrip vir datawetenskaplikes se aandrang op 'n meer objektiewe wetenskapsbeoefening wat op data eerder as op vooropgestelde idees gegrond is – veral in die geesteswetenskappe.

'n Groot hoeveelheid kritiek is reeds uitgespreek teenoor grootdatabenaderings, veral vanuit die geesteswetenskappe, en tot 'n groot mate teenoor Anderson se idee dat data vir sigself 'spreek'.[21] Kritiek let gewoonlik daarop dat data – afkomstig van die Latyns 'datum' wat "gegewe" beteken – nie bloot gegewe is nie, maar geskep word (Schöf 2013:3-4; Puschmann & Burgess 2014:1691-1694; Boyd & Crawford 2012:667-668). Kritiese argumente teen grootdata sluit veral 'n bevraagtekening in of grootdata werklik so objektief is as wat beweer word wanneer die data eerstens deur 'n mens geskep word en verder deur 'n mens geïnterpreteer word. Die data spreek nie vanself nie; die ontleding dui eerder op die navorser se eie ingesteldheid en word ook beïnvloed deur die manier waarop die datastel saamgestel is (Van Dijck 2014:201-202). Die navorser se wetenskapsfilosofiese en metodologiese voorkeure, sosiaal-sielkundige eienskappe, ideologie en wetenskaplike paradigma – kortweg, sy menslikheid – beïnvloed alles die navorsingsproses (Mouton & Marais 1990:10-12), en dit verander nie in 'n era van grootdata nie. Grootdata ontsnap nie van die menslike nie, alhoewel voorstaanders beweer dat dit wel die geval is. Mahrt en Scharkow (2013:30) skryf in weerwil van Anderson dat die gebruik van grootdata steeds teorie en 'n wetenskaplike navorsingsmetodologie benodig, en Omand, Bartlett, en Miller (2012:822) voer aan dat grootdata baat kan vind by die in-diepte kennis wat reeds binne die akademie bestaan.

In 'n sekere opsig is Anderson se stelling egter nie só vergesog nie. Gegronde teorie is juis 'n teoretiese raamwerk waar die navorser deur sy data gelei word (Charmaz 2014; Byrne & Callaghan 2014:199), en die induktiewe metode het 'n lang geskiedenis in die wetenskap (Mouton & Marais 1990:113; Reichertz 2004:303). 'n Bekende voorbeeld van 'n induktiewe navorsingstrategie is Stanley Milgram se bogenoemde 'klein-wêreld'-studie (1967). Milgram het koeverte aan ewekansig geselekteerde individue in Kansas en Nebraska gestuur en hulle gevra om dit na 'n gespesifiseerde ontvanger in Boston aan te stuur. Hy het gevra dat hulle, indien hulle nie die persoon ken nie, die koevert aan iemand gee of stuur wat hulle glo dit by die ontvanger sou kon uitkry. Alhoewel

21 Sien in hierdie verband onder andere Faltesek (2013:409), Bollier (2010:5-7), Boyd en Crawford (2012:666) en Kitchin (2014:5).

die meerderheid van koeverte weggeraak het, het 'n paar tog hul bestemming bereik, en Milgram het uitgewerk dat dit slegs 'n gemiddeld van ses stappe geneem het vir die koevert om by die teiken uit te kom. Só is die afleiding rakende ses grade van verwydering, dat almal slegs ses stappe van mekaar verwyder is, gemaak. Watts en Strogatz (1998) het dieselfde beginsel geneem en toegepas op die internasionale filmakteurnetwerk – as 'n voorbeeld van 'n sosiale netwerk – en bevind dat akteurs slegs 'n gemiddeld van 3,65 stappe van mekaar verwyder is (soos hierbo vermeld). Hulle het ook ander netwerke ondersoek, met soortgelyke resultate. In beide gevalle is die klein-wêreld-teorie vanuit die eksperiment gegenereer: in Milgram se geval slegs ten opsigte van sosiale netwerke, en in Watts en Strogatz se geval ten opsigte van 'n aantal komplekse netwerke. Hieruit is die voorspelling gemaak dat enige twee rolspelers binne enige komplekse netwerk gemiddeld met 'n klein aantal stappe met mekaar verbind kan word – die teorie is gegenereer vanuit die eksperiment. Grootdatabenaderings se klem op die induktiewe metode is dus geensins 'n totale nuwe epistemologiese benaderingswyse nie, alhoewel dit gereeld so gesien word.

Ook is grootdatabenaderings nie só vry van teorie soos wat voorgegee word nie. Kitchin (2014:4) skryf dat die datagedrewe benadering van grootdata nie teorie veronderstel nie, en haal die sagtewaremaatskappy Ayasdi aan ter stawing van hierdie siening. Wat Kitchin egter nie merk nie is dat Ayasdi se eie inligtingsvideos beklemtoon dat hul fondasie in wiskundige grafiekteorie lê, onder andere deur verwysings na Leonard Euler, wat as die vader van wiskundige grafiekteorie en die oorsprong van die netwerkteorie beskou word (Senekal 2014b:11-12). Daar is dus nie alleen 'n lang geskiedenis van die induktiewe metode in die wetenskap nie, maar sagtewaremaatskappye soos Ayasdi se ontledingsbeginsels is wel gegrond in wetenskaplike teorieë (dieselfde geld byvoorbeeld vir Palantir en Starlight VIS, wat ook in die netwerkteorie geanker is). Die skuld vir die siening dat grootdata 'n teorievrye benaderingswyse is val op datawetenskaplikes wat beweer dat hulle die data toelaat om vanself te spreek terwyl dit nie in die praktyk die geval is nie.

'n Mens sou dus kon sê dat grootdatabenaderings in weerwil van Anderson *nie* die einde van teorie beteken nie, maar eerder 'n groter klem op die induktiewe metode plaas. Hiermee saam integreer grootdatabenaderings met die sisteem- en netwerkteorie, soos blyk uit rekenaarprogrammatuur soos Palantir, Starlight VIS en Ayasdi se toepassings van hierdie beginsels (sien verderaan).

2.5 Die einde van die kenner

Grootdata het 'n nuwe soort wetenskaplike geskep: die datawetenskaplike. Dié term is deur Jeff Hammerbacher en D.J. Patil geskep toe hulle onderskeidelik by Facebook en LinkedIn gewerk het (Davenport 2014:92; Krishnan 2013:255) en dui op 'n wetenskaplike wat groot hoeveelhede data, wat in 'n verskeidenheid formate voorkom, op 'n innoverende wyse ontleed. Mayer-Schönberger en Cukier (2013:134-145) stel voor dat grootdata se klem op inligtingstegnologie kan beteken dat rekenaargeletterde navorsers uiteindelik kenners sal vervang, soos wat algoritmes die kritici op Amazon vervang het. Volgens hierdie outeurs (2013:142) is die kenner juis 'n simptoom van 'n tyd toe daar nie met grootdatastelle omgegaan kon word nie, soos die geval is met steekproefneming. Die toekoms van die wetenskap lê aldus die outeurs in rekenaargeletterdheid, of meer spesifiek in die vermoë om grootdatastelle te ontleed, eerder as dissipline-spesifieke kennis.[22] McAfee en Brynjolfsson (2012:65) verwys na sulke kenners as HIPPO's – die "Highest-Paid Person's Opinion" – en alhoewel die outeurs spesifiek na die besigheidsektor verwys, word sulke kenners ook in die akademie gevind. Uitstaande professors verdien natuurlik heelwat meer as junior lektors, en hul sienings en aanbevelings dra meer gewig. In 'n datagedrewe wêreld verskuif die klem na dié wat die data kan versamel en ontleed, en die rol van 'n 'seekoei' verskuif vanaf 'n kenneropinie oor antwoorde na die formulering van vrae: waarna behoort datawetenskaplikes te kyk? (McAfee & Brynjolfsson 2012:66). Kenners ('seekoeie') het dus steeds 'n plek, maar hul funksie word eerder om hul kennis te gebruik om navorsing te rig.

In 'n grootdatawêreld behoort daar dus 'n simbiose geskep te word tussen kenners en datawetenskaplikes, sodat eersgenoemde voorstelle kan maak rakende wat om te ondersoek, en laasgenoemde die grootskaalse ondersoek kan behartig. 'n Voorbeeld uit eie ondervinding waar so 'n benadering gevolg is, is Senekal en J.-A. Stemmet (2014) se studie van Jamie Uys se rol in die Afrikaanse filmbedryf as netwerk. Stemmet het reeds oor baie jare in-diepte navorsing oor Jamie Uys onderneem, en sy deskundigheid en ervaring is geïntegreer met Senekal se tegniese en teoretiese agtergrond. Sodoende kon 'n nuwe benaderingswyse gevolg word wat baie breër na Uys se posisie in die

22 Agrawal et al. (2011:7) skryf egter dat geen verantwoordelike navorser alles aan 'n rekenaar sal toevertrou nie; die navorser se eie begrip en kritiese ontledingsvaardighede bly steeds noodsaaklik om sin te maak uit rekenaarmatige ontledings (sien ook McAfee en Brynjolfsson (2012:66)). Volgens Davenport (2014:110) beteken tegnologie ook nie dat die mens vervang word nie, maar die mens se rol verander wel. Hoe meer kompleks die ontledings word, hoe meer datawetenskaplikes word benodig om die ontledings te behartig.

filmbedryf kyk, saam met 'n kenner wat die probleemstellings van die studie kundig kon stuur. Só 'n simbiose het ook die voordeel dat dit die sterkpunte van beide navorsers se kennissisteme maksimaliseer.

Hierdie verskuiwing bring ook mee dat datawetenskaplikes noodwendig interdissiplinêr te werk gaan (Loukides 2010:8), en Hitzler en Janowicz (2013:233) stel voor dat die probleme en geleenthede betreffende grootdata ook interdissiplinêr aangespreek word. In hierdie opsig stem grootdatabenaderings weer eens ooreen met die holistiese, interdissiplinêre benaderings wat deur Von Bertalanffy bepleit is en veral gerealiseer word binne die teorieë van kompleksiteit, sisteme en netwerke (Von Bertalanffy 1972:416; Johnson 2009:18; Bar-Yam 1997:1). De Beer (2003:124) beaam dat vakkundige werk dissiplinêre grense oorsteek. Volgens hom word vakkundige werk nie gebind of beperk deur sulke grense nie, maar trek dit lyne na ander dissiplines, vestig verbindings, en soek nimmereindigend na kennis.[23] In hierdie opsig bring grootdatabenaderings nuwe moontlikhede vir interdissiplinêre samewerking in die wetenskap. Let ook op die bronne in die huidige boek: daar is publikasies van so ver en uitlopend as die linguistiek, fisika, rekenaarwetenskap, literatuurwetenskap, militêre intelligensie en sosiologie – alles interdissiplinêre navorsing wat met behulp van inligtingstegnologie onderneem is.

Oorsese universiteite, waaronder Columbia en die Universiteit van New York, het dan ook onlangs begin om gespesialiseerde datawetenskaplikes op te lei (Provost & Fawcett 2013:57; Davenport 2014:102-103). In Suid-Afrika bied die Noordwes-Universiteit se Potchefstroomkampus 'n BCom graad in Ekonomie en Informatika aan wat grootdata-ontledings en programmeringsvaardighede insluit. Ook in die geesteswetenskappe is daar 'n toenemende besef van die belangrikheid van inligtingstegnologie, en Jockers (2013:13) skryf dat kursusse wat spesifiek toegespits is op die digitale geesteswetenskappe reeds by Stanford, Kings Kollege in Londen, die Nasionale Ierse Universiteit in Maynooth, University College in Londen, Trinity College in Dublin en verskeie universiteite in Kanada geskep is.

'n Verdere algemene kritiek wat teen grootdata uitgespreek word en wat aansluit by die vorige bespreking, is dat dit die digitale gaping vergroot (Abreu & Acker 2013:550; Boyd & Crawford 2012:673-675; McNeely & Hahm 2014:308). Eerstens is daar die vraag wie toegang kry tot die data: groot maatskappye soos Facebook, Amazon en Google beskik oor die data, maar stel dit nie geredelik vry nie, en veral nie aan

23 Sien ook Wilden (1980:241).

navorsers binne die akademie nie. Dit baat die akademiese navorser byvoorbeeld min dat selfoonmaatskappye groot hoeveelhede data rakende selfoonoproepe genereer en berg as dit nie aan die navorser beskikbaar is nie. Barabási was in 'n bevoorregte posisie om hierdie data te ontleed, aangesien sulke samewerking nie gereeld plaasvind nie. 'n Digitale gaping ontstaan tussen die besigheidsektor en die akademie, waar eersgenoemde meer data tot sy beskikking het.

Tweedens is daar die vraag oor die ontleding van data, wat Manovich (2012) die 'data-ontledingskloof' noem: vaardighede en rekenaarprogrammatuur kom ter sprake wat ontleders in die akademie (en veral in ontwikkelende lande) uitsluit. Hare (2014:73) skryf dat grootdata selfs in die besigheidswêreld gewoonlik net gebruik word deur dié met die vaardighede en/of groot hoeveelhede geld. Starlight VIS en Palantir was vir die outeurs van hierdie boek te duur om aan te skaf, terwyl dit wel gebruik word deur die VSA se intelligensiedienste.[24] Tableau spreek hierdie probleem aan deur 'n gratis weergawe aan studente te verskaf (juis om ook ontledingsvaardighede te ontwikkel), en Actian het 'n gratis gemeenskapsweergawe, maar heelwat ander sagtewaremaatskappye het nie soortgelyke ontwikkelingsinisiatiewe nie. Sommige grootdataprogrammatuur soos Hadoop, MapReduce, R, en die programmeringstale Python, Hive en Pig – wat van die belangrikste tegnologiese hulpmiddels in grootdata-ontledings is – is wel gratis beskikbaar, maar verg 'n deeglike kennis van hierdie programmatuur asook rekenaarprogrammeringsvaardighede (Davenport 2014:132). Dié met die kennis en vaardighede word opgeraap deur groot besighede, aangesien daar tans 'n groot tekort aan datawetenskaplikes bestaan, wat beteken dat dié vaardighede uit die akademie stroom.

Nie alleen gaan die akademie oor die algemeen mank aan sulke vaardighede nie, maar ook ontbreek die nodige tegniese vaardighede binne die geesteswetenskappe, wat beteken dat studies van aspekte van die menslike beweegruimte deur datawetenskaplikes vanuit die natuurwetenskappe, wat nie noodwendig 'n opleiding in die geesteswetenskappe het nie, onderneem word. Die natuurwetenskappe het 'n lang geskiedenis in die aanwending van inligtingstegnologie vir navorsingsdoeleindes, en veral die fisika is tans die dominante studierigting as dit kom by die opleiding van datawetenskaplikes (Davenport 2014:91). Fisici soos Duncan Watts, Albert-Lázló Barabási en Mark Newman wend hulle gereeld tot die bestudering van sosiale fenomene. Hierdie natuurwetenskaplikes kom vanuit 'n inligtingstegnologie-georiënteerde agtergrond waar tegnologiese vaardighede reeds vroeg in hul loopbane

24 Die outeurs het kwotasies van onderskeidelik $55 000 en $147 000 vir hierdie programmatuur bekom.

vasgelê is; die geesteswetenskaplike navorser het nie hierdie begronding nie, en sukkel dus om hierdie tegnieke aan te wend. Barabási (2011:15) skryf dat die fisika die terrein van kompleksiteitsteorie reeds 'n geruime tyd lank oorheers, maar dat ontwikkelings in rekenaarwetenskap veroorsaak het dat fisika geen kompetisie meer het nie. Die gevaar bestaan dat fisici en ander datawetenskaplikes uiteindelik ook die geesteswetenskappe sal domineer en die gesaghebbende kenners op dié terrein word (wat Barabási beweer reeds gebeur het).

Tinati et al. (2014:665) let daarop dat alhoewel dit mag voorkom asof die beskikbaarheid van grootdata die studie van sosiale fenomene wegskuif van sosiologie na rekenaarwetenskap en fisika, die klem in sulke studies op die identifisering van netwerkpatrone eerder as op 'n verkenning van die veld self val. Dieselfde geld vir film, waar die groot hoeveelheid studies oor die internasionale filmakteurnetwerk eerder fokus op die netwerkeienskappe van hierdie industrie as op die generering van kennis rakende die industrie self. Daar is dus tans steeds 'n plek vir die geesteswetenskappe, omdat data-gebaseerde navorsingsmetodes op 'n ander manier na die menslike beweegruimte kyk, maar interdissiplinêre samewerking kan daartoe bydra dat navorsers tot ryker en dieper insigte in die menslike beweegruimte kan kom, en die geesteswetenskappe se gesag-hebbendheid oor sosiale fenomene help behou.

Die ontleding van grootdata bevorder dus die gaping tussen besigheid en die regering aan die een kant, en die akademie aan die ander kant, en kan ook die gaping tussen die natuur- en geesteswetenskappe vergroot. Grootdatabenaderings se klem op interdissiplinêre samewerking werk hierdie gaping tot 'n mate teë, maar dit is belangrik dat toenemende samewerking aangemoedig en geïnisieer word.

2.6 Die einde van reduksionisme

Alhoewel Mayer-Schönberger en Cukier nie spesifiek skryf oor die teoretiese skuif wat grootdata veronderstel nie (omdat hulle glo dat grootdata die einde van teorie aankondig), behoort dit uit die voorafgaande duidelik te wees dat teorie steeds 'n belangrike komponent van die wetenskap in die era van grootdata uitmaak. In een opsig impliseer grootdata wel die einde van sekere teorieë en wetenskapsfilosofiese sienings: die omvattendheid van datastelle en die fokus op die bestudering van datastelle in die geheel beteken die einde van *reduksionisme*. Reduksionisme haal 'n studie-objek uitmekaar, bestudeer onderdele in isolasie, en verteenwoordig die tradisionele wetenskaplike metode (Galitski 2012:52). Reduksionisme veronderstel dat die geheel die somtotaal van sy onderliggende elemente is, en daarom sal 'n beter begrip van die werking van

elemente ook tot 'n beter begrip van die funksionering van die geheel lei. Buchanan (2003:72) skryf byvoorbeeld dat 'n motorwerktuigkundige na 'n foutiewe part sal soek wanneer daar 'n probleem met 'n motor is, en dat die identifisering van die foutiewe part dan sal bydra daartoe dat die probleem opgelos word. In 'n menslike konteks sou daar ook vanuit hierdie siening aangevoer kon word dat 'n beter begrip van individue se persoonlikhede, houdings en sienswyses sal bydra daartoe dat 'n gemeenskap as geheel beter verstaan kan word, byvoorbeeld die siening dat Duitse motors van 'n goeie gehalte is omdat Duitsers baie noukeurig te werk gaan.

Von Bertalanffy (1950:134, 1968:18-19, 1972:411) het reeds sy kritiek uitgespreek teenoor wetenskaplike metodes wat verskynsels uitmekaar haal en weer aanmekaar sit in 'n poging om tot 'n beter begrip van die werklikheid te kom, en navorsers soos Bar-Yam (1997:11), Plsek (2001:311), Barabási (2011:14) en Luke en Stamatakis (2012:358) het later ook hierdie tendens gekritiseer. Aangesien die geheel in 'n oop of komplekse sisteem meer as die somtotaal van die onderdele is, is dit belangrik om nie slegs die onderdele te bestudeer wanneer 'n mens tot 'n beter begrip van die geheel wil kom nie (Von Bertalanffy 1972:411). Volgens Barabási (2011:14) het ons die limiete bereik van wat deur middel van reduksionisme bepaal kan word, want alhoewel reduksionisme deur die twintigste eeu tot belangrike en bruikbare insigte gelei het, is die samehang van elemente in die wêreld 'n belangrike faset wat ook in ag geneem moet word indien die wetenskap tot 'n begrip van meer komplekse fenomene wil kom.

Grootdata is ingebed in wat Watts (2004:14) die "Connected Age" noem: let byvoorbeeld op hoeveel van bogenoemde voorbeelde te make het met sosiale media, die web, en die verhoudinge tussen entiteite. Barabási (2003:7) skryf dat die mens in die hedendaagse wêreld al hoe meer bewus geword het daarvan dat niks in isolasie plaasvind nie, en dat die wetenskap daarom 'n groter klem op verhoudinge lê as vantevore.[25] Dié klemverskuiwing van die deel na die geheel is volgens Barabási (2009:413) die direkte gevolg van die beskikbaarheid van groter digitale datastelle, sowel as rekenaarprogrammatuur wat hierdie groter datastelle kan ontleed, en as sodanig is die teorie van komplekse netwerke stewig ingebed in grootdatabenaderings. In 'n poging om die komplekse hedendaagse wêreld te begryp, het die teorie van kompleksiteit onlangs opgang begin maak in 'n groot verskeidenheid dissiplines, maar soos Barabási (2011:15) skryf is kompleksiteitsteorie meesal gegrond op simulasies.[26] Hierteenoor is

25 Sien ook Costa et al. (2011:331).

26 Sien ook Byrne en Callaghan (2014:40).

die teorie van komplekse netwerke die produk van die induktiewe metode en altyd verankerd in data. Barabási let daarop dat rekords van menslike handeling reeds in verskeie databasisse gestoor word: e-pos- en telefoonrekords dokumenteer ons sosiale en professionele interaksies, reisrekords en GPS-navigasiestelsels vang ons reispatrone en fisiese bewegings op, en kredietkaartmaatskappye hou rekords van ons inkopies en vermaakgewoontes. Hoewel hierdie datastelle volgens hom in die verkeerde hande Orwelliaanse gereedskap van mag verteenwoordig, bied hulle vir wetenskaplikes ongelooflike insig in menslike gedrag. Kombineer hierdie vermoë om data te versamel met die gesofistikeerde instrument van die netwerkteorie, wat verhoudinge tussen miljoene individue kan ontleed, en jy kry 'n blik op 'n ongekende geleentheid om menslike dinamika te kwantifiseer (Barabási 2005a:639).

Holistiese benaderings het natuurlik 'n lang geskiedenis binne die wetenskap, met Ludwig von Bertalanffy (1950, 1968) en Kurt Lewin (1951) van die bekendste wetenskaplikes in die 20ste eeu om voor te stel dat fenomene bestudeer moet word binne die komplekse web van interaksies waarbinne hul funksioneer. In hierdie opsig is grootdatabenaderings se klem op die geheel en interafhanklikheid dus geensins nuut nie, soos kompleksiteitsteorie óók nie 'nuut' is nie, maar voortbou op die algemene sisteemteorie van Von Bertalanffy (Schneider & Somers 2006). Binne die antropologie is daar reeds 'n lang geskiedenis van die toepassing van sisteem- en netwerkteorie, onder andere deur mense soos Lewin (1951), Bavelas (1948) en Nadel (1957), terwyl die sosiologie deur onder andere Moreno (1934) en Freeman (2004) ook 'n lang geskiedenis van netwerk- en sisteemteoretiese benaderings het. In politieke wetenskap is Kilcullen (2010, 2013) se onlangse beskrywings van konfliksisteme as *komplekse sisteme* veral van belang aangesien hy tans as een van die voorste kenners van terroris-strategie en teenopstand bekendstaan, en in die sielkunde word die sisteemteorie al vir baie jare toegepas (Vorster 2003). Die Afrikaanse letterkunde is sedert die tagtigerjare deeglik bewus daarvan dat tekste nie in isolasie funksioneer nie, maar ingebed is in 'n komplekse netwerk van verhoudinge tot ander tekste en rolspelers binne die literêre sisteem, sowel as ekstraliterêre sisteme soos die politiek, ekonomie en sosiale strukture (Viljoen 1986; Senekal 1987). Senekal (1987:34, 44) skryf: "Die literêre werk word nie as 'n outonome eenheid gesien nie, maar as die produk van die literatuursisteem", en "[k]ultuur is 'n web, nie 'n stukkie drukwerk in isolasie nie".

Wat van die grootdatabenadering is dan op 'n teoretiese vlak nuut? Eerstens breek grootdatabenaderings weg van reduksionisme, wat beteken dat grootdata meesal veronderstel dat die wetenskap beoefen word teen die agtergrond van die netwerk- en/of

sisteemteorie. Grootdata maak dus nie gebruik van reduksionistiese navorsingspraktyke wat fenomene dekontekstualiseer en uitmekaarhaal in 'n poging om tot 'n beter begrip van daardie fenomeen se funksionering te kom nie. Dié benadering is wel voorgestel deur verskeie teoretici, onder andere Von Bertalanffy, maar grootdata beklemtoon die geheel en verskaf die middele om die geheel mee te bestudeer. Laasgenoemde is veral belangrik omdat die geheel tegniese uitdagings aan die navorser bied wat eers onlangs met behulp van rekenaarprogrammatuur die hoof gebied kon word: om *al* die inhoud van 'n literatuursisteem in berekening te bring is byvoorbeeld nie moontlik sonder rekenaarprogrammatuur nie. Tweedens breek grootdata weg van kompleksiteitsteorie se klem op vooropgestelde teorieë (lees deduksie), en verskuif die klem na induksie en die gepaardgaande datagedrewe netwerkteorie. Grootdata werk altyd vanaf die data na die teorie, en maak nie voorsiening vir ongegronde simulasies en modelle nie. Watts en Strogatz (1998) se klein-wêreld-model, sowel as Barabási en Albert (1999) se skaalvrye model, is beide modelle wat gegenereer is vanuit data, nooit andersom nie.

Kortom beteken dit dat grootdatabenaderings in die voetspore van teoretici soos Von Bertalanffy volg deur die geheel en verhoudinge tussen elemente te beklemtoon, die middele verskaf om die geheel mee te bestudeer, en die induktiewe metode vooropstel. Grootdatabenaderings staan so teenoor reduksionisme.

2.7 Gevolgtrekking

Soos McNeely en Hahm (2014:309) aanvoer, is grootdata deurspek met potensiaal en probleme. Die debat oor die bruikbaarheid van grootdata is maar in 'n beginfase, en sal in die toekoms verder gevoer moet word om te bepaal of dit werklik so bruikbaar is soos voorstaanders aanvoer. Wat egter nie ontken kan word nie, is dat grootdata iets is waarmee die geesteswetenskappe rekening sal moet hou – dit kan nie geïgnoreer word nie.

Onses insiens kan die geesteswetenskappe baat vind by Schöf (2013:9) se voorstel dat die gaping tussen klein- en grootdata oorbrug moet word deur groter 'slim' data en 'slimmer' grootdata te benut – 'n siening wat ook implisiet is in Agrawal et al. (2011) se bespreking van die potensiaal en probleme van grootdata. Met 'slim' data bedoel Schöf (2013:3) gestruktureerde en semi-gestruktureerde, akkurate data. Hy (2013:9) wys egter daarop dat ten einde die groot/klein gaping te oorbrug, nuwe metodes toegepas sal moet word, byvoorbeeld om van rekenaarprogrammatuur of spanne gebruik te maak. Amazon se Mechanical Turk is 'n voorbeeld van laasgenoemde, waar groot hoeveelhede mense data aanlyn 'skoonmaak', met ander woorde onakkuraathede uit die weg ruim, en dit teen

'n klein vergoeding (Loukides 2010:5; Watts 2011:48-50). Die konsep is vergelykbaar met die aanstelling van 'n groot aantal navorsingsassistente wat elk 'n klein komponent van die werk verrig, en gaan (soos in die geval van navorsingsassistente) gepaard met die risiko dat sommiges foute sal maak. Hierdie beginsel word onder andere aangewend om karaktererkenning te verbeter deur die CAPTCHA's (Completely Automated Public Turing test to tell Computers and Humans Apart) wat 'n mens gebruik om inligting op webblaaie in te vul (Agrawal et al. 2011:11-12); die karakters wat 'n mens moet oortik is afkomstig vanuit 'n teks waar karaktererkenningsprogrammatuur onseker was oor wat die letter is. Alhoewel die oënskynlike funksie van die CAPTCHA is om seker te maak dat dit 'n mens is wat inligting op die webblad invul en nie 'n robot nie, dien dit ook die doel om karaktererkenning te verbeter. So werk ons almal saam om die akkuraatheid van karaktererkenning te verbeter en die inligting wat in ouer dokumente vasgevang is te ontsluit.

'n Mens sou ook rekenaarprogrammatuur kon inspan om datastelle skoon te maak, byvoorbeeld Google se OpenRefine, wat 'n mens help om data in gestruktureerde kolomme te sorteer. Andersins kan rekenaarprogrammatuur aangewend word om so veel as moontlik 'donkiewerk' uit die verwerkingsfase van 'n navorsingsprojek te verwyder en daardeur tyd te skep om data skoon te maak – dié benadering is gevolg om beide die datastel rakende die Afrikaans filmindustrie as dié oor die Afrikaanse letterkunde saam te stel. Sodoende kan groter datastelle saamgestel word, wat wel nie só groot is soos dié wat Walmart gebruik nie, maar steeds aansienlik groter is as wat tot op hede saamgestel kon word en op sigself omvattende datastelle verteenwoordig.

'n Verdere manier om die gaping tussen groot en klein data deur middel van 'n middeweg te oorbrug, is deur rekenaarprogrammatuur in te span wat 'n verskeidenheid formate kan hanteer en groter datastelle kan hanteer. Om hierdie rede fokus 'n latere hoofstuk op die rekenaarprogram NVivo, wat 'n manier verskaf om kwalitatiewe navorsing te herposisioneer in 'n grootdata-opset deur die ryk verskeidenheid formate te akkommodeer wat met grootdata gepaard gaan én kwalitatiewe navorsing met behulp van groter datastelle te onderneem. Eers moet daar egter aandag gewy word aan waar groter datastelle bekom kan word, wat die onderwerp van die volgende hoofstuk is.

Hoofstuk 3

Dataversameling in 'n era van grootdata

Die versameling van bruikbare datastelle is 'n belangrike komponent van die navorsingsproses en figureer dus ook in elke bespreking van navorsingsmetodologie (Du Toit & Smith-Muller 2003:135; Mouton & Marais 1990:25; Vermeulen, Lategan & Litheko 2011:15). In 'n grootdata-omgewing is die versameling van bruikbare data egter meesal digitaal, en soos voorheen vermeld kan dit 'n beduidende hindernis in die navorsingsproses wees wanneer die navorser 'n slukkie water vanuit 'n tsunami probeer drink. Indien die navorser waarde wil put uit die groot hoeveelhede inligting wat wêreldwyd beskikbaar is, moet hy van inligtingstegnologie gebruikmaak. Die gevorderde ontledingsmetodes wat in die komende hoofstukke bespreek word, is slegs van nut indien die navorser oor data beskik om te ontleed. In hierdie hoofstuk word veral gefokus op databronne wat vir die geesteswetenskappe van belang is. Let egter daarop dat die huidige hoofstuk relatief konserwatiewe dataversameling bespreek; 'n latere hoofstuk bespreek meer radikale strategieë.

Bose (2008:516) onderskei tussen passiewe versameling, wat deurlopende inligtingsbehoeftes ondersteun, en aktiewe versameling, wat meer doelgerig te werk gaan, soos vervolgens meer breedvoerig bespreek word.[27]

3.1 Passiewe versameling

Bose (2008:518) skryf dat daar soms na die passiewe versameling van inligting verwys word as "information push" (vergelykbaar met wat McKee, Koltutsky en Vaska (2009:3) "current awareness alerting" noem), wat behels dat inligting 'n organisasie vryelik binnestroom, terwyl die aktiewe versameling van inligting die intrek van inligting in die inligtingstelsel van 'n organisasie behels. Passiewe versameling behels die opstel van die nodige infrastruktuur, byvoorbeeld die intekening op relevante nuusbriewe, of die opstel van rekenaarprogrammatuur om die web outomaties te monitor vir nuwe inligting en dit dan in die navorser se databasis in te trek. Dit het primêr die voordeel dat 'n wye verskeidenheid onderwerpe gedek kan word, in teenstelling met aktiewe

27 Punte 3.1 en 3.2 is gebaseer op Senekal (2012a).

versameling, wat gewoonlik geskied in antwoord op 'n spesifieke aanvraag en doelgerigte inligting oplewer. Soos die geval by die Nasionale Afrikaanse Letterkundige Museum en Navorsingsentrum (NALN), en die Universiteit van die Vrystaat se SA Media en die Suid-Afrikaanse Taalregte Monitor (SALRM), is die opbou en instandhouding van knipselversamelings 'n passiewe versamelingsaktiwiteit (in Bose se sin van die woord): koerante, tydskrifte en joernale word deurgegaan vir relevante artikels, geïndekseer en in die databasis gestoor.

Wat digitale passiewe versameling aanbetref is daar heelwat rekenaarprogrammatuur wat die gebruiker in staat stel om webblaaie outomaties te monitor vir hersiene inligting, asook om outomatiese internet-soektogte uit te voer. Dit beteken dat die navorser outomaties in kennis gestel word wanneer nuwe inligting oor 'n onderwerp beskikbaar is. Sommige van dié programmatuur is gratis, terwyl ander 'n inskryffooi vereis:

- ChangeDetect (www.changedetect.com)
- MetaProducts Offline Explorer (www.metaproducts.com)
- Check&Get (activeurls.com)
- HTTrack (www.httrack.com)

Ander programmatuur wat spesifiek vir die akademiese milieu ontwerp is, is RefAware en IngentaConnect, wat die web deurlopend monitor vir nuwe inligting oor 'n gegewe onderwerp, en die navorser in kennis stel van nuwe inligting wat gevind word. Sulke programme maak dit vir die navorser maklik om op hoogte te bly van die nuutste ontwikkelings in sy veld. Ander gratis opsies is ticTocs en Google Alerts. Deur bloot in te teken op relevante elektroniese nuusbriewe (byvoorbeeld die van Stratfor of LitNet) ontvang die navorser ook gereelde inligting wat hom in kennis stel van ontwikkelings in sy veld. Selfs Facebook kan van waarde wees vir die navorser: deur koerante en ander organisasies se Facebook-blaaie te volg, word nuwe inligting outomaties aan die navorser deurgegee. Hierdeur kan die navorser op hoogte bly van wat in die wêreld aangaan.

Die passiewe versameling van inligting het primêr die voordeel dat 'n wye verskeidenheid onderwerpe gedek kan word, in teenstelling met aktiewe versameling, wat beteken dat nuwe terreine makliker ontdek word. Terselfdertyd is daar egter ook aansienlik meer geruis (inligting wat nie relevant is tot die onderhawige studie nie) in verhouding tot relevante inligting.

3.2 Aktiewe versameling

Die versameling van betroubare inligting binne 'n koste-effektiewe tydraamwerk is een van die sleutels tot suksesvolle navorsing. Daar word soms hierna verwys as 'inligtingherwinning', dit wil sê die interdissiplinêre wetenskap wat die soek na dokumente, inligting binne dokumente en dokumente se metadata in databasisse en op die wêreldwye web insluit (Moisil 2009:25). Pirolli en Card (1999:11) beklemtoon dat die aktiewe versameling van inligting gedryf word deur die noodsaak om die soektog so koste-effektief as moontlik af te handel. 'n Mens kan inderwaarheid aan die inligtingsoeker dink as 'n 'inligtingsroofdier' wat ten doel het om die 'inligtingsprooi' só te kies dat die verhouding tussen insette en 'voedingswaarde' gemaksimaliseer word.

Pirolli en Card (2005:3) onderskei tussen die versameling van inligting wat van onder na bo gedryf word (met ander woorde van data tot gevolgtrekkings), en versameling wat van bo na onder gedryf word (waar die soektog deur 'n bepaalde hipotese gedryf word). In eersgenoemde doen die navorser 'n soektog na relevante databasisse, doen navraag, ensovoorts, en versamel dan data in 'n digitale biblioteek vir verdere ontleding. Wanneer deur hierdie dokumente gelees word, word daar uiteraard telkens nog inligting bekom, wat die navorser dan weer noodsaak om terug te keer na die versamelingsfase wanneer nuwe bronne ontdek of teoretiese benaderings teëgekom word. Dan word 'n aktiewe soektog op die web of databasisse soos EBSCOhost gedoen deur sleutelwoorde, outeurs, publikasies of die titels van artikels te gebruik. 'n Soektog van bo na onder kan plaasvind wanneer bestaande opvattinge bevraagteken of bevestig moet word (Pirolli & Card 2005:4).

Ongeag of 'n soektog van onder na bo of van bo na onder gedryf word, aktiewe soektogte behels die raadpleging van veral drie virtuele terreine wat in die Inligtingsera relevant is vir die geesteswetenskappe: webblaaie, digitale databasisse, en sosiale media.

3.2.1 Die web

In 1958, kort na die Sowjetunie die eerste mens in die buitenste ruim ingestuur het, het die Amerikaanse president, Dwight D. Eisenhower, DARPA gestig. Die oorspronklike doelwit van DARPA was om tegnologiese verassings soos die lansering van Sputnik, wat gewys het dat die Sowjets die VSA na die ruimte voorgespring het, te verhoed (Defense Advanced Research Projects Agency 2005:1). Hoewel dié missie later grotendeels oorgeneem is deur NASA (National Aeronautics and Space Administration), het DARPA sedertdien vele tegnologiese mylpale bereik, insluitend die ontwikkeling van

sluipbomwerpers soos die F 117 Nighthawk, onbemande lugvaartuie soos die Predator en Global Hawk, en die globale posisioneringstelsel (GPS). DARPA se bekendste mylpaal is egter die ontwikkeling van wat vandag bekend staan as die internet. Die konsep van die internet is soortgelyk aan Paul Baran van RAND (Research And Development) Corporation se siening dat die VSA 'n kommunikasienetwerk moes stig wat nie deur 'n kernaanval van die Sowjetunie vernietig sou kon word nie (Barabási 2003:143 e.v.; Caldarelli 2013:186). Baran het in 1964 voorgestel dat 'n verspreide struktuur met 'n groot aantal oorbodige skakels die beste weerstand teen so 'n aanval sou kon bied (Baran 1964). Sy voorstelle is om verskeie redes geïgnoreer, maar DARPA het met 'n soortgelyke ontwerp vorendag gekom (Barabási 2003:145). Die internet is toe in 1969 as 'n militêre netwerkstelsel gestig, en kort daarna was daar 'n koppeling tussen vier rekenaars by die Universiteit van Kalifornië in Los Angeles, die Universiteit van Kalifornië in Santa Barbara, die Universiteit van Stanford, en die Universiteit van Utah (Dolowitz, Buckler & Sweeney 2008:1). Teen 1972 was daar negentien rekenaars in die VSA met mekaar verbind (Buchanan 2003:76). Toegang was egter beperk tot 'n paar honderd rekenaars binne die akademiese gemeenskap, en wat ons vandag as die internet ken het eers gedurende die 1980's ontwikkel.

In 1989 het die Engelse fisikus Tim Berners-Lee en die Belgiese rekenaarwetenskaplike Robert Cailliau by CERN (die Europese organisasie vir kernnavorsing) voorgestel dat 'n mens 'n web van bladsye, wat met skakels verbind is, sou kon gebruik om inligting op die internet te stoor en te navigeer, en hulle het dit die wêreldwye web genoem (Caldarelli 2013:199-200). Wat algemeen bekendstaan as die internet is in werklikheid die versameling rekenaars en bedieners – die fisiese skakels – wat die hardeware-komponent van die wêreldwye web vorm (Newman 2010:18-28). Daarenteen is die wêreldwye web die kuberruimte – die netwerk van webblaaie en hiperskakels waarmee 'n mens van dag tot dag te doen kry (Newman 2010:63-67). Teen Junie 1993 was daar om en by 130 sulke webblaaie in die wêreld, en dit het eksponensieel gegroei – teen Junie 1998 was daar 2 410 067 webblaaie, en teen 2003 was daar 40 936 076 (Lima 2011:56). Tans is daar só baie webblaaie op die web as geheel (dit wil sê die oppervlak- en diep-komponente) dat die getal onbekend is (Olcott 2012:110),[28] maar daar word geskat dat dit uit biljoene webblaaie bestaan. Google het onlangs reeds 'n triljoen unieke webadresse (URL of

28 Deel van die probleem om die aantal webblaaie te tel is natuurlik dat webblaaie deurlopend verander, en selfs vir spesifieke gebruikers geskep word, byvoorbeeld 'n "Wishlist" op Amazon of Kalahari. 'n Mens sou ook persone se Facebook-profiele kon noem as 'n webblad wat voortdurend verander.

Uniform Resource Locator) (Craig & Ludloff 2011:4; Appel 2011:11) en die inhoud van 50 biljoen webblaaie (Shroff 2013:9) geïndekseer.

Baie van die groot aanlyn maatskappye is in die laat negentigerjare gestig. eBay en Amazon het beide in 1995 begin, en Google het in 1998 in die voetspore van Lycos en Yahoo! gevolg om die gebruiker in staat te stel om inligting op die groeiende web op te spoor (Craig & Ludloff 2011:3). Dié internetmaatskappye was van meet af die toonaangewendes as dit by die bestuur van grootdata kom, en vandag is Google steeds die toonbeeld van wat grootdata kan vermag. Van rekenaarmatige vertalings deur Google Translate wat op astronomiese korpusse en komplekse algoritmes steun, tot die selfbesturende kar wat gebruikmaak van sensors en geografiese data, tot die identifisering van epidemies deur korrelasies tussen mense se soekterme te ontgin, is Google by uitstek 'n grootdatamaatskappy (Davenport 2014, Mayer-Schönberger & Cukier 2013).

Die wêreldwye web is sedert die vroeë 1990's geheel en al vry van 'n sentrale beheerliggaam. Terwyl akademiese publikasies die goedkeuring van hekwagters soos redakteurs en keurders moet kry voor dit gepubliseer word, kan enigeen iets op die web plaas. Die gebrek aan kontrole lei natuurlik daartoe dat baie inhoud op die web van 'n swak gehalte is, maar 'n mens moet egter versigtig wees om alle webblaaie oor dieselfde kam te skeer: baie van dieselfde inligting wat in druk verskyn, verskyn ook op die web (byvoorbeeld akademiese artikels), en gesaghebbende wetenskaplikes publiseer ook hul insigte op webblaaie. Die kommer wat tans oor die kwaliteit van inligting wat op die internet beskikbaar is uitgespreek word, is vergelykbaar met die kommer wat in Europa uitgespreek is oor die kwaliteit van gedrukte materiaal toe die drukpers vir die eerste keer algemeen in gebruik geneem is. Bawden en Robinson (2009:182) skryf dat vrae oor die waarheid en betroubaarheid van wat in skrif aangebied word mense nog altyd bekommer het, van die propagandapamflette van die sewentiende-eeuse oorloë en godsdienstwis tot die webkameras van die politici van die hedendaagse leefwêreld. Wikipedia – ten spyte van akademiese wantroue – is grotendeels akkuraat: 'n studie van Giles (2005) het bevind dat die gemiddelde artikel op Wikipedia bykans so akkuraat is as die gemiddelde artikel in *Encyclopedia Britannica*.[29] Foute kom wel voor, maar dit geld ook vir die gedrukte media. Een artikel in 'n ISI-gelyste akademiese joernaal verwys byvoorbeeld na "James Woods" se roman, *Fields of fire*, terwyl die outeur in werklikheid James Webb is. 'n MA-verhandeling by 'n Suid-Afrikaanse universiteit beweer ook dat Tony Buckingham die private militêre maatskappy, Executive Outcomes, in 1993

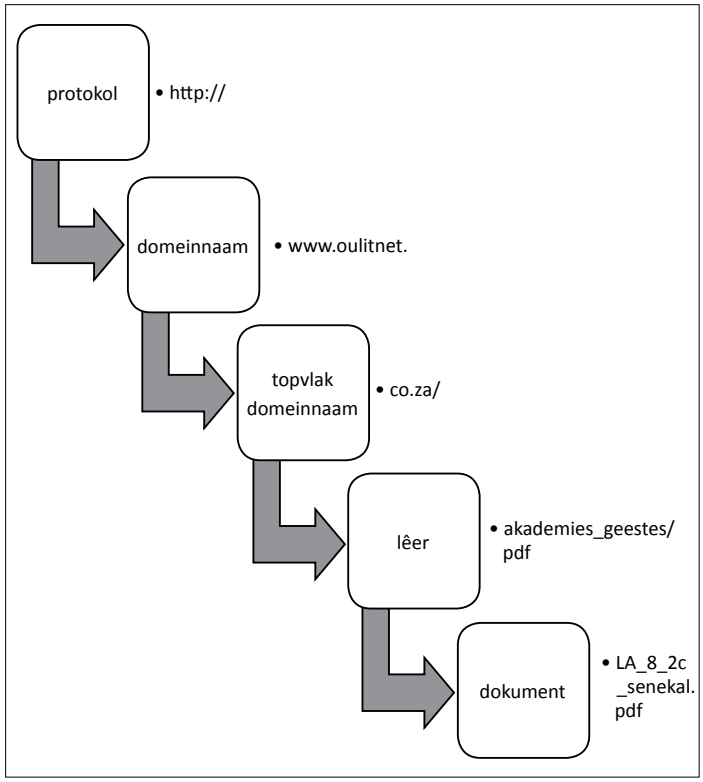
29 Sien ook Olcott (2012:112).

gestig het, terwyl dit eerder Eeben Barlow was wat dié maatskappy in 1989 gestig het. Daar kom soms beduidende feitefoute in geakkrediteerde akademiese joernale, MA-verhandelings en gepubliseerde boeke voor, en dieselfde geld vir inligting wat op die web beskikbaar is.

Alhoewel dit nie 'n waarborg vir akkuraatheid is nie, is 'n goeie riglyn om soveel as moontlik bronne te raadpleeg – hiervoor is die elektroniese opsporing en ontleding van bronmateriaal weer eens deurslaggewend. Meer gesaghebbende bronne kan geraadpleeg word, wat beteken dat navorsing op 'n stewiger basis gegrond is. Die tyd wat spandeer word om 'n enkele bron in 'n biblioteek te vind, kan gebruik word om talle bronne aanlyn op te spoor, en sodoende kan die navorser sy feite meer omvattend, en vinniger, kontroleer.

Dít beteken egter geensins dat 'n biblioteek (of die gedrukte boek) nutteloos is nie. Biblioteke verleen elektroniese toegang tot joernale wat digitaal beskikbaar is, en daarsonder sou die navorser se toegang tot publikasies uiters beperk wees. Afgesien van hierdie toegang is daar steeds plek vir die gedrukte boek, hetsy aangekoop of uitgeneem, aangesien daar juis so baie bruikbare inligting in boeke opgesluit lê. 'n Mens kan ook elektroniese weergawes van boeke gebruik, soos Kindle of EPUB, en dit dan saam met artikels in 'n digitale biblioteek stoor (deur middel van byvoorbeeld Qiqqa). Sodoende kan die navorser ook digitaal met sy boeke omgaan. Manning Publications het in hierdie opsig 'n publikasiemodel wat navorsing in die 21ste eeu bevorder: wanneer 'n mens 'n boek in hardekopie aankoop, verskaf hulle ook 'n digitale kopie in EPUB-, Kindle- en PDF-formaat, wat beteken dat 'n mens 'n digitale kopie benewens 'n hardekopie kan berg.

Figuur 7 is 'n grafiese voorstelling van die samestelling van 'n internetadres. Wanneer daar kennis gedra word van die samestelling van die internetadres kan daar sodoende addisionele inligting op die internet opgespoor word. Die voorbeeld in Figuur 7 toon die adres van 'n PDF-dokument. Soms sal 'n spesifieke soektog so 'n dokument vind, maar ander dokumente op dieselfde webwerf kan ook van waarde wees, selfs al het 'n Google-soektog dit nie gevind nie (moontlik omdat die soekterme dit nie ingesluit het nie). Om sodanige dokumente te vind, kan die spesifieke dokumentnaam in die adresstaaf verwyder word – in hierdie geval sal die lêer 'akademies_geestes' dan gevind word, wat na *LitNet Akademies (Geesteswetenskappe)* verwys. Die [/] skei die verskillende vlakke van mekaar, dus kan telkens met een vlak opgegaan word wanneer alles na die [/] verwyder word.



Figuur 7. 'n Grafiese voorstelling van 'n internetadres

Die web bestaan uit die oppervlakweb, die bladsye wat deur soekenjins geïndekseer kan word, en die diepweb, wat gewoonlik spesiale toegang vereis (byvoorbeeld aanlyn databasisse waar 'n gebruikersnaam en wagwoord benodig word). Dit is belangrik om hiervan kennis te neem, aangesien die onvermoë van soekenjins om toegang te verkry tot die diepweb beteken dat 'n soektog deur middel van 'n soekenjin slegs deur die oppervlakweb soek (Appel 2011:15).

Google is die voorste soekenjin in die Westerse wêreld (Noruzi 2005:171; Ripple 2006:98; Olcott 2012:107), soveel so dat die naam teen 2003 die status van 'n werkwoord verkry het. Google is egter nie die enigste soekenjin nie, en dié maatskappy se markaandeel wissel van 98% in Litaue tot slegs 3% in Suid-Korea; oor die algemeen domineer Baidu in China, Naver in Suid-Korea, en Yandex in Rusland (Olcott 2012:107). Die soekenjin wat gebruik word, bepaal tot 'n groot mate watter inligting gevind word, soos Olcott (2012:108-110) illustreer met verwysing na Google, Bing, Yahoo!, Yandex en Baidu.

Hoewel daar ander geskikte soekenjins vir dié doel is, is die wetenskaplike been van Google, Google Scholar (www.scholar.google.co.za), een van die geskikste soekenjins vir navorsingsdoeleindes. Dit is ontwikkel deur die Indiese rekenaarwetenskaplike Anurag Acharya en fokus op akademiese publikasies, hetsy joernale, webblaaie of boeke. Google Scholar ontleed outomaties watter artikels die meeste as verwysings dien in bibliografieë, wat dit maklik maak vir die navorser om te bepaal wie die leiers op 'n bepaalde terrein is (Noruzi 2005:171). Dit kan ook met vrug gebruik word deur die naam van die outeur of titel van 'n bekende artikel in te tik en dan te kyk watter ander artikels hierna verwys. Die datum van publikasie kan gespesifiseer word om meer onlangse navorsing om te spoor. Noruzi (2005:173) noem ook dat Google Scholar interdissiplinêre navorsing aanmoedig, wat – soos vroeër bespreek – veral belangrik is in die hedendaagse wêreld.

Google Scholar het onder andere die volgende voordele (Noruzi 2005:174):

- Dit verskaf internasionale toegang tot akademiese publikasies.
- Dit laat navorsers toe om breë, omvattende, en multidissiplinêre soektogte te loods wat ooreenkomste tussen dissiplines uitlig.
- Daar is geen vooroordeel teenoor vakke nie (maar wel sprake van 'n taalvooroordeel).
- Google Scholar is nie beperk tot artikels nie – tegniese verslae, verhandelings, proefskrifte en akademiese PowerPoint-voorleggings tel ook onder die resultate.
- Navorsers kan soektogte doen volgens sleutelwoorde, outeurs of titels.
- Navorsers kan op een slag artikels soek wat oor vele jare gepubliseer is.

Gehanno, Rollin, en Darmoni (2013) het bevind dat Google Scholar voldoende is vir 'n literatuuroorsig, maar Boeker, Vach, en Motschall (2013) bevraagteken op hul beurt of dit werklik deeglik genoeg is. Alhoewel daar nie eenstemmigheid is oor of Google Scholar genoegsaam gepas is nie, is dit 'n belangrike soekenjin vir navorsingsdoeleindes.

3.2.2 Databasisse

Google is nie altyd die beste plek om te begin soek nie: die volume soekresultate waarmee die navorser gekonfronteer word kan oorweldigend wees. Daarby dra 'n mens in die beginfase van 'n projek nie altyd genoeg kennis van 'n onderwerp om die regte soekterme te gebruik nie. Die term 'netwerk' word in verskeie kontekste en velde gebruik, maar as 'n mens byvoorbeeld op soek is na die sosiologiese toepassing daarvan, wat dikwels binne sosiale netwerk analise (SNA) voorkom, kan die navorser dit spesifiseer. Die navorser beskik egter nie noodwendig oor dié kennis wanneer hy 'n nuwe terrein aandurf nie, en daarom kan dit sinvol wees om eers gespesialiseerde databasisse te raadpleeg.

Gespesialiseerde databasisse kan ook inligting oplewer wat nie deur Google op die sogenaamde diepweb gevind kan word nie. Wanneer 'n soekenjin inligting soek, soek dit nie werklik deur alles wat beskikbaar is op die internet nie, maar eerder deur 'n indeks daarvan (Dolowitz, Buckler & Sweeney 2008:62). Om hierdie rede word die oorgrote meerderheid van relevante inligting dikwels oor die hoof gesien, en volgens Olcott (2012:110) is die diepweb tussen 100 en 1000 keer groter as die oppervlakweb. Aansluitend hierby is dit ook beter om 'n databasis soos Sabinet te gebruik wanneer daar spesifiek na Suid-Afrikaanse bronne gesoek word, omdat hierdie bronne nie altyd in 'n gewone Google Scholar-soektog opduik nie. Die bekendste en nuttigste databasisse binne die geesteswetenskappe sluit in:

- JSTOR (www.jstor.org)
- EBSCOhost (www.ebscohost.com)
- Sabinet (www.sabinet.co.za)

JSTOR is in 1995 gestig om internasionale toegang tot wetenskaplike publikasies te verleen. Dit beskik onder andere oor die volle uitgawes van meer as 2 000 akademiese joernale – wat neerkom op meer as 6 miljoen artikels (JSTOR 2013). Benewens 'n besondere groot reikwydte oor dissiplines, sluit dit ook ouer artikels in, wat van groot waarde kan wees vir 'n historiese ondersoek.

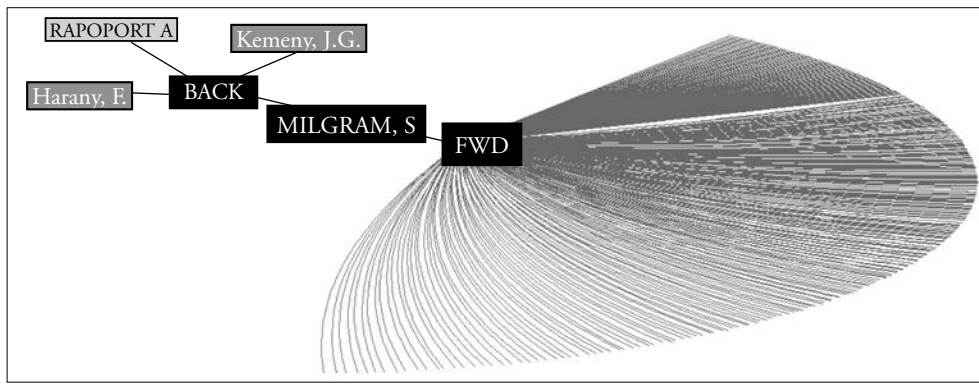
EBSCOhost beskik ook oor historiese rekords in digitale formaat, maar anders as JSTOR verleen dit ook toegang tot versamelings primêre bronne, insluitend briewe, administratiewe rekords en dagboeke. Dié inligtingsdiens het toegang tot 375 digitale databasisse, wat op hul beurt toegang verleen tot 420 000 digitale boeke en 355 000 joernale.

Teenoor dié versamelings lyk Sabinet se SA ePublications, wat toegang tot 300 joernale bied, wel klein, maar SA ePublications bied 'n diens wat van spesiale belang is vir Suid-Afrikaanse navorsers. Eerstens verskyn nie alle materiaal wat deur SA ePublications beskikbaar is in algemene soekresultate nie, wat beteken dat belangrike plaaslike navorsing oor die hoof gesien kan word as hierdie databasis nie geraadpleeg word nie. Tweedens kan dit 'n meer hanteerbare wegspringplek wees, veral omdat die fokus op die plaaslike val; soektogte op dié databasis lewer plaaslike resultate, en nie die miljoene potensiële resultate wat internasionale dienste oplewer nie.

Daar bestaan ook 'n wye verskeidenheid ander databasisse wat nie noodwendig toegang tot akademiese publikasies verskaf nie, maar dalk eerder tot primêre bronne. 'n Belangrike versameling in hierdie verband is SA Media se knipseldatabasis, wat

knipsels oor 'n groot verskeidenheid onderwerpe in 'n digitale formaat beskikbaar stel. Die databasis strek sover terug as 1978, wat baie nuttig is vir historiese ondersoeke. Daar is byvoorbeeld 53 151 knipsels oor apartheid; 3 461 oor huursoldate; en 16 432 oor Jamie Uys.[30] Die navorser kan hierdie knipsels aflaai en in sy digitale biblioteek berg vir verdere gebruik. Regeringsorganisasies kan ook van hulp wees, byvoorbeeld vir amptelike dokumentasie of statistieke. Statistiek Suid-Afrika (www.statssa.gov.za) is hier van besondere belang, asook die databasis van die Suid-Afrikaanse Polisiediens. Elke dissipline het toegang tot dergelike bronne wat gespesialiseerde soektogte moontlik maak, en dit is 'n goeie idee om hierdie webwerwe op te spoor en te merk as gunstelinge, sodat daar later maklik daarheen teruggekeer kan word.

'n Spesiale geval wat hier genoem kan word, is Thompson Reuters se Web of Science (www.thomsonreuters.com/thomson-reuters-web-of-science/). Die voordeel van dié databasis lê daarin dat dit op verwysings gebaseer is, wat beteken dat die bronne van artikels, asook watter artikels die oorspronklike artikel aanhaal, maklik gevind kan word. Die navorser kan dan moeiteloos deur die verwysingsnetwerk navigeer met die wete dat die meeste bronne waarna hy kyk waarskynlik relevant vir sy projek is. Die grafiese voorstelling in Figuur 8 toon die verwysingsnetwerk van Milgram se "The small world problem" (1967), met die bronne waarna Milgram verwys aan die linkerkant, en die artikels waarin daar na sý artikel verwys word aan die regterkant.



Figuur 8. Milgram (1967) se verwysingsnetwerk

30 Soos op 19 Februarie 2014. SA Media se bestaande databasis word vanaf 2015 deur Sabinet oorgeneem en nuwe knipsels word nie meer bygevoeg nie. Die voorbeelde wat hier verskaf is, het betrekking op vorige studies wat ek (Senekal) onderneem het.

Die digte verwysingsnetwerk aan die regterkant dui daarop dat hierdie 'n seminale artikel is: Web of Science dui aan dat daar 450 keer daarna verwys is. Aan die linkerkant is Milgram se bronne (slegs drie). Die navorser kan van hier af na volteksartikels gaan, aangesien Web of Science 'n aanduiding gee van wat om verder te lees. Dit mag dalk insiggewend wees om Milgram se bronne te raadpleeg, maar die digte verwysingsnetwerk aan die regterkant verteenwoordig nuwer artikels wat verwys na hierdie seminale publikasie, en die feit dat hierdie artikels na Milgram verwys is 'n aanduiding daarvan dat dit ook vir die navorser relevant mag wees.

3.2.3 Sosiale media

Sosiale media het oor die afgelope dekade 'n belangrike bron van inligting geword. Aangesien daar geen keuringsproses is wat die inhoud van sosiale media bepaal nie, is die feitelikheid daarvan altyd verdag, maar nietemin gee dit 'n aanduiding van die kwessies wat bespreek word en die menings wat mense ten opsigte van brandende vraagstukke huldig. Wat veral belangrik is, is hóé feite weergegee word, eerder as die feite self – veral wanneer 'n mens 'n idee wil kry van die emosies wat rondom 'n saak heers. Omand, Bartlett en Miller (2012:806) skryf byvoorbeeld dat die polisie sentiment kan monitor, wat mag dui daarop dat geweld binnekort gaan uitbreek, en in so 'n geval maak dit min saak of dié wat inligting op sosiale media plaas 'reg' of 'verkeerd' is; wat tel is wat gesê word. Op 'n soortgelyke manier sou 'n navorser persepsies kon bestudeer soos dit uitgebeeld word op sosiale media. Daar is egter ook gevalle waar sosiale media die hoofstroommedia vooruit was in die oordrag van nuus: die Wikipedia-inskrywing oor die bomaanvalle in Londen op 7 Julie 2005 en die Twitter-berig oor die dood van Michael Jackson val as voorbeelde by (Olcott 2012:112). Dan is daar ook die reeks tweets van die IT-konsultant Sohaib Athar wat vanaf kort voor middernag op 1 Mei 2011 vanuit Abbottabad, Pakistan, berig het oor helikopters in die gebied – Butcher (2011) noem hom "the guy who unwittingly live-tweeted the raid on Bin Laden". Sosiale media het dus 'n rol om te speel in die oordrag van nuus én as bronmateriaal vir ander studies, en daarom verskaf hierdie afdeling agtergrond oor dié fenomeen.

Die blog is 'n digitale persoonlike weergawe van gebeure, en het rondom 1997 ontstaan (Olcott 2012:84). Daar bestaan natuurlik verskeie soorte blogs, van persoonlike dagboeke tot kritiek op regeringsbeleid, wat dit 'n primêre bron (soos briewe en onderhoude) van 'n individu se sienings maak. Die feit dat inligting op 'n blog verskyn, maak dit nie noodwendig verdag nie: Joan Hambidge plaas kritiese kunsbesprekings op haar blog, en hierdie kan op dieselfde manier as ander resensies gebruik word. Indien die

navorser spesifiek blogs wil deursoek, kan Google se gespesialiseerde soekfunksie, Blog Search (www.google.com/blogsearch), gebruik word.

Die eerste video is in 2005 op YouTube (www.youtube.com) gelaai, en volgens die webwerf word daar tans[31] elke uur 'n 100 ure se videomateriaal – van absurde video's van mense wat die lirieke van liedjies verkeerd interpreteer tot kursusmateriaal – daarop gelaai. YouTube kan op allerlei maniere vir navorsingsdoeleindes aangewend word: daar is video's oor hoe om rekenaarprogrammatuur te gebruik, nuusberigte, onderhoude met, byvoorbeeld, skrywers, dokumentêre en dies meer. Let ook op die kommentaar onder 'n video (en die antwoorde op dié kommentaar), wat dikwels 'n aanduiding is van die menings wat mense huldig jeens die onderwerp wat in die video ter sprake kom.

Sosiale media word algemeen geassosieer met Facebook (wat in 2004 geloods is), maar daar bestaan ook ander platforms wat van waarde kan wees, soos Friendster (gestig 2002) en Myspace (gestig 2003). In Rusland domineer Livejournal (gestig 1999), terwyl Hi5 (gestig 2003) gewild is in Nepal, Mongolië, Thailand, Roemenië, Jamaika, Sentraal-Afrika, Portugal en Latyns-Amerika. In China is Renren (in 2005 gestig as die Xiaonei Network) gewild, terwyl Cyworld (gestig 1999) die markleier in Suid-Korea is (Olcott 2012:85).

Twitter is in 2006 gestig, en gee gebruikers die geleentheid om ander te 'volg'. Sedert Mei 2014 is daar ook 'n Afrikaanse sosiale media platform wat baie soortgelyk aan Twitter is, Toeter (www.toeter.co.za), wat op dieselfde beginsel werk.

Gespesialiseerde soekenjins wat op sosiale media fokus, soos in detail deur Bazzell (2013) bespreek, sluit in:

- Tweet Archivist (www.tweetarchivist.com)
- Twitwheel (www.twitwheel.com)
- TweetReach (tweetreach.com)
- Twiangulate (twiangulate.com)
- Addict-o-matic (addictomatic.com)
- Socialmention (www.socialmention.com)
- Backtweets (backtweets.com)
- Convoflow (convoflow.com)
- IceRocket (www.icerocket.com)
- Topsy (topsy.com)
- Samepoint (samepoint.com)

31 Soos op 11 Julie 2014.

3.3 Gevolgtrekking

Wanneer die navorser sy bronmateriaal (akademiese artikels sowel as primêre bronne) opgespoor het, kan dit in 'n digitale biblioteek gestoor word. Qiqqa, wat juis so 'n digitale biblioteek is, is in 2009 deur James Jardine by die Universiteit van Cambridge ontwikkel, en is spesifiek geskep as 'n PDF-bestuurprogram binne die akademiese navorsingskonteks. Smith (2012) skryf dat Qiqqa in die eerste plek 'n PDF-leser is, wat annotasie, kodering, notas, soekfunksies en kruisverwysings ondersteun, en ook vele handige addisionele funksies besit, soos 'n dinkskrum, karaktererkenning (wat 'n mens toelaat om deur PDF's te soek), en 'n omvattende liasseringstelsel, sodat daar nie deur talle artikels gesoek hoef te word om 'n enkele stuk inligting te vind nie. Wanneer die navorser só 'n stuk gereedskap gebruik, spaar hy tyd deur 'donkiewerk' soos die liassering van artikels en die latere opspoor van inligting daarin uit te skakel. Dié tydsbesparing, sowel as die tydsbesparing wat deur bogenoemde soektogte teweeggebring word, is noodsaaklik in 'n era van grootdata, want in die volgende fases van die navorsingsproses (verwerking en ontleding) sal 'n groot hoeveelheid tyd belê moet word om om te gaan met die data self. Aangesien die meeste navorsers binne die geesteswetenskappe in Suid-Afrika kwalitatief georiënteerd is, word dataverwerking by data-ontleding geïntegreer in die volgende hoofstuk, wat fokus op die herposisionering van kwalitatiewe navorsing deur middel van programmatuur soos NVivo. Die kwantitatief georiënteerde navorser kan gereedskap soos LDA (Latent Dirichlet Allocation) (Blei, Ng & Jordan 2003; Blei et al. 2004; Griffiths & Steyvers 2004) of MALLET (McCallum 2002) aanwend om data in 'n meer geskikte formaat te verwerk.

Rekenaargesteunde kwalitatiewe data-ontledingsprogrammatuur (RGKDOP): 'n Herposisionering van kwalitatiewe navorsingsmetodes

4.1 Inleiding

Rekenaargesteunde kwalitatiewe data-ontledingsprogrammatuur (RGKDOP)[32] is reeds sedert die tagtigerjare beskikbaar om kwalitatiewe data-ontleding te ondersteun. Sedert die algemene gebruik van die wêreldwye web in die negentigerjare het daar op hierdie gebied beduidende ontwikkeling plaasgevind in 'n poging om die kwalitatiewe navorser in staat te stel om die inligtingsontploffing tegemoet te gaan, en ook juis op so 'n manier dat kwalitatiewe navorsing bly voortbestaan in 'n wêreld wat al hoe meer oorrompel word deur kwantitatiewe navorsingsmetodes. Daar bestaan wel 'n wye verskeidenheid RGKDOP – waarvan ATLAS.ti en MAXQDA van die vernaamste voorbeelde is – maar hierdie hoofstuk fokus op een van die leiers op die gebied van RGKDOP, naamlik NVivo, in 'n poging om die herposisionering van kwalitatiewe navorsing binne 'n era van grootdata te bespreek. Alhoewel NVivo gewoonlik weggelaat word in besprekings van grootdata (dit word byvoorbeeld nie in Mayer-Schönberger en Cukier (2013) of Davenport (2014) genoem nie), is dit onses insiens 'n belangrike program wat die uitdagings van grootdata aanspreek en dit boonop binne die raamwerk van kwalitatiewe navorsing doen, wat steeds die dominante navorsingsmetode in heelwat velde binne die geesteswetenskappe is.

4.2 NVivo en grootdata

'n Artikel deur Tom Richards, in 2002 gepubliseer in die *International Journal of Social Research Methodology*, bied 'n interessante intellektuele geskiedenis van twee vorme van RGKDOP, naamlik NUD*IST (Non-Numerical Unstructured Data Indexing

32 Computer-Assisted Qualitative Data Analysis Software (CAQDAS) in Engels.

Searching and Theorizing) en NVivo (laasgenoemde is 'n verbeterde weergawe van eersgenoemde en is in 1999 bekendgestel). Terwyl NUD*IST, waarvan Richards by La Trobe Universiteit in Melbourne mede-ontwikkelaar in 1981 was, die maatstaf vir kwalitatiewe navorsingsprogrammatuur was, is NVivo in staat daartoe om ontledings te vermag wat sy voorloper nie kon nie. Eerstens, in teenstelling met NUD*IST, is NVivo in staat om karaktergebaseerde kodering te akkommodeer, wat beteken dat eenhede teks nie vooraf gespesifiseer hoef te word nie. Tweedens is NVivo daartoe in staat om om te gaan met teks in verskillende kleure, lettertipes, groottes en style, iets wat NUD*IST nie kon doen nie (Richards 2002:208). Derdens laat NVivo navorsers toe om veranderings aan te bring soos hulle kodeer, 'n vermoë wat nog nie in soortgelyke programmatuur van die 1980's bestaan het nie (Richards 2002:208). NVivo is dus die erfgenaam van NUD*IST, maar is spesifiek ontwikkel vir die dinamiese hedendaagse Inligtingsera.

4.2.1 NVivo en volume

Soos voorheen vermeld bied die grootte en verskeidenheid van data in 'n grootdatawêreld 'n beduidende uitdaging aan navorsers. NVivo kan dié uitdagings aanspreek, eerstens deur die navorser toe te laat om akkuraat met groter hoeveelhede data om te gaan. Richards (2009:33) skryf dat dit maklik is om kwalitatiewe data te skep, maar wat egter nie so maklik is nie, is om die data te organiseer en te bestuur. Boonop is kwalitatiewe data dikwels lomp en neem groot hoeveelhede ruimte in beslag, veral wanneer klank, video en hoë resolusie beelde ter sprake is. Die standaard weergawe van NVivo kan projekte so groot as 10 gigagrepe hanteer, en NVivo Server kan projekte van onbeperkte grootte hanteer. Wanneer individuele dokumente bó 20 megagrepe elk is, word dit outomaties ekstern geberg (met ander woorde nie in die projek self nie), en dié limiet kan ook aangepas word na 100 megagrepe elk. Dié oplossing verminder dan die grootte van die projek self, wat beteken dat daar met 'n groot hoeveelheid data omgegaan kan word, en 'n hele datastel kan binne 'n enkele projek geakkommodeer word (die grootte van die datastel word uiteindelik slegs beperk deur die hoeveelheid digitale stoorplek wat beskikbaar is). NVivo se vermoë om met groot datastelle om te gaan maak dit veral bruikbaar in 'n grootdatawêreld, juis omdat die navorser nie verplig word om met slegs 'n deel van die data te werk nie. Silverman (2013:269) identifiseer onder andere die volgende voordele in die gebruik van rekenaarprogrammatuur soos NVivo vir kwalitatiewe navorsingsdoeleindes:

1. Spoed in die hantering van groot volumes data, wat die navorser bevry om verskeie analitiese vrae te verken, asook om groter monsters te kan ontleed;

2. verbetering van akkuraatheid, insluitend die bemiddeling van die kwantifisering van verskynsels en die identifisering van afwykende gevalle;
3. fasilitering van samewerking, insluitend die ontwikkeling van konsekwente koderingskemas.

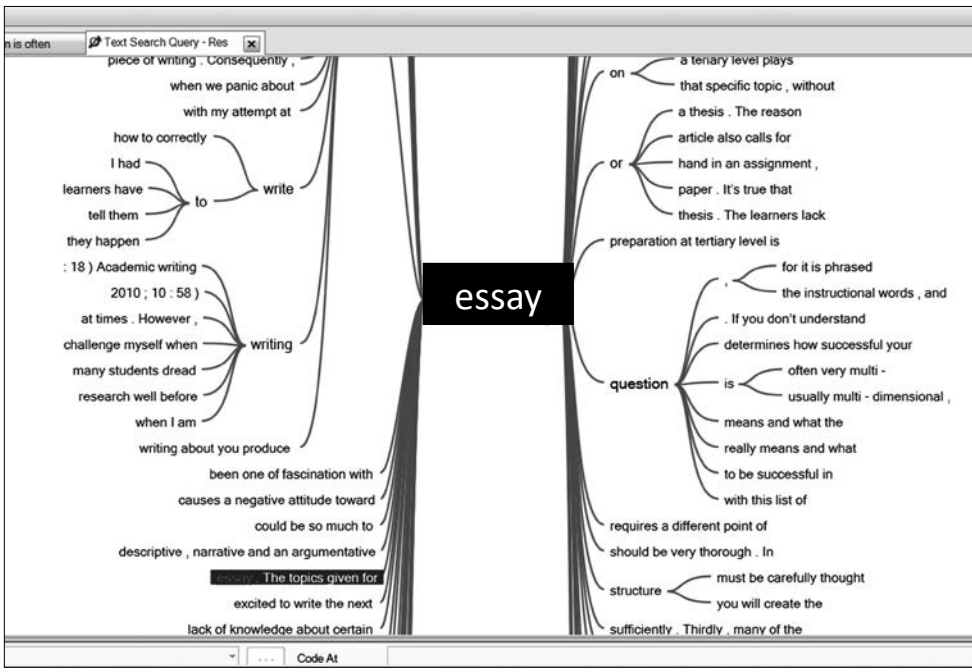
Wat die bestuur van groot datastelle aanbetref het NVivo ook 'n hele aantal maniere om vrae van sy datastel te vra, byvoorbeeld soektogte na kernwoorde, gebruiksfrekwensies, en die outomatiese groepering van dokumente op grond van hoe soortgelyk hulle is. Dié funksies stel die navorser in staat daartoe om inligting te herwin, en veral om ontdekkings te maak wat nie deur 'n eenvoudige sleutelwoordsoektog uitgelig word nie. NVivo stel ook kwalitatiewe navorsers in staat om hul data deur middel van visualiseringstegnieke soos modelle, grafieke of kaarte grafies voor te stel (Bazeley & Jackson 2013:3). Soos later bespreek word, is die visualisering van data 'n belangrike komponent van die ontleding van groot datastelle, en NVivo kan dus hier ook met vrug benut word.

NVivo stel die navorser in staat om met sy datastel as geheel om te gaan en konsepte te ontleed binne die netwerk van verhoudinge waarbinne dit funksioneer. 'n Soektog na gebruiksfrekwensies sal byvoorbeeld deur die hele korpus soek, en dieselfde gebeur wanneer groeperings van dokumente gedoen word. Kyk byvoorbeeld na die gebruiksfrekwensie van die woord "essay" wat in Figuur 9 uitgelig word.

Al die verskillende verbindings waarin dié woord voorkom kan in so 'n visualisering gesien word, en in so 'n voorstelling is die woord gesitueer binne die verhoudinge waarin dit funksioneer. Die konteks waarin 'n woord of tema voorkom kan ook gespesifiseer word om 'n breër of kleiner konteks aan te dui. Dié gekontekstualiseerde uitbeelding is 'n belangrike voordeel van rekenaarprogrammatuur soos NVivo, aangesien dit altyd belangrik is om te let op die individuele datapunt se verhouding tot ander datapunte.

NVivo het wel 'n beperking wat die omgang met grootdata aanbetref – alhoewel die grootte van die projek nie 'n begrensing is nie, is die program se onvermoë om berekenings oor verskillende verwerkers te versprei wel problematies. Wanneer datastelle petagreepgroottes begin aanneem is berging nie die enigste probleem nie; berekening kan ook baie lank neem om uit te voer. Berekenings in NVivo is oor die algemeen nie só vinnig soos byvoorbeeld in Tableau nie, wat probleme sal veroorsaak as 'n mens met baie groot datastelle van honderde duisende of miljoene dokumente werk. Dít is moontlik 'n rede waarom NVivo gewoonlik weggelaat word in besprekings van grootdata. Onses insiens behoort dié beperking egter nie probleme te veroorsaak vir die oorgrote meerderheid navorsers binne die geesteswetenskappe nie, aangesien die datastelle wat

deur 'n akademiese navorser ontleed word gewoonlik nie sulke astronomiese groottes aanneem nie. Vir besigheid en militêre intelligensie is dié egter 'n verdoemende probleem.



Figuur 9. Gebruiksfrekwensies in die konteks van 'n hele korpus

4.2.2 NVivo en verskeidenheid

NVivo toon inderdaad die vermoë om ongestruktureerde data soos dagboekinskrywings, fokusgroepkommentaar, video's, boeke en tydskrifte te ontleed en daaruit sin te maak (Wiedemann 2013). Die program se ontledings is nie beperk tot gestruktureerde data soos in die formaat van Microsoft Excel of Access nie. Dit het 'n beduidende voordeel vir die geesteswetenskappe, aangesien ongestruktureerde data deurslaggewend is vir kwalitatiewe navorsing, en die meerderheid bronne in die geesteswetenskappe is ook gewoonlik in 'n ongestruktureerde formaat (byvoorbeeld koerantuitknipsels). 'n Instrument soos NVivo kan gebruik word om ongestruktureerde data te organiseer en te bestuur, wat dit makliker maak vir navorsers om sin van die deurmekaarspul te maak. Meer spesifiek stel NVivo volgens Bazeley en Jackson die navorser in staat om

te organiseer en rekords te hou van die slordige bronmateriaal wat deel uitmaak van 'n kwalitatiewe navorsingsprojek. Dit sluit nie net die rou datalêers van onderhoude, vraelyste, fokusgroepe of waarnemings in nie, maar ook gepubliseerde navorsing, beelde, diagramme, klank, video, webbladsye, ander dokumentêre bronne, rowwe notas en idees wat in memorandums neergeskryf is, inligting oor databronne en konseptuele kaarte van wat aangaan in die data (Bazeley & Jackson 2013:3).

Bykomend tot dié formate stel NVivo die navorser in staat om met sosiale media om te gaan, en NVivo werk ook saam met SurveyMonkey om oop vraelyste te hanteer.

Een van die grootste voordele van NVivo is dat dit nie nodig is om ongestruktureerde data eers in 'n gestruktureerde formaat om te skakel nie, aangesien data ontleed kan word in die formaat waarin dit bestaan. Dit bring onder andere mee dat dié rekenaarprogrammatuur nie die verwerkingsfase van die navorsingsprojek vergroot nie (anders as byvoorbeeld Tableau), maar natuurlik kan NVivo slegs spesifieke ontledingsaksies vermag en is dit nie so kragtig soos Tableau of R wat statistiese ontleding aanbetref nie.

4.3 Kritiek op die gebruik van RGKDOP

Daar is reeds vroeg kritiek uitgespreek teen die gebruik van rekenaarprogrammatuur vir ontledingsdoeleindes binne kwalitatiewe navorsing. Redes wat gereeld aangevoer word sluit in tegniese probleme (Rademaker, Grace & Curda 2012:3), vrae oor die geldigheid van bevindinge, en die totstandkoming van 'n ontledingsafstand tussen die navorser en sy data (Deakin, Wakefield & Gregorius 2012:605). Desnieteenstaande is daar, veral sedert 2000 (Gibbs 2014:278), 'n groot aantal navorsers binne die geesteswetenskappe wat gebruik maak van kwalitatiewe data-ontledingsprogrammatuur om hul data te bestuur (Meyers, Bennett & Lysaght 2004; Ferguson 2010, Uzum 2010, Veletsianos, Kimmons & French 2013; Brokensha & Greyling 2014). Die kritiek en oplossings rakende RGKDOP word in die hieropvolgende afdelings bespreek.

4.3.1 Die programmatuur word die metode

Tien jaar gelede het MacMillan en Koenig (2004:179) opgemerk dat navorsers binne die geesteswetenskappe onkrities met hul rekenaarprogrammatuur omgaan, en die afwesigheid van kritiese besinnings oor rekenaarprogrammatuur is steeds 'n probleem. 'n Ondersoek na 40 studies – met een uitsondering (Hickson 2012) – wat oor die afgelope vyf jaar gepubliseer is, meesal binne rekenaargesteunde kommunikasie of RGK, en wat gebruikmaak van NVivo dui daarop dat daar steeds geen kritiese besinning is oor

die gebruik van dié gereedskap nie. Opmerkings is beperk tot ’n blote noem van ’n paar beweerde voordele soos gesien in Tabel 5.

Tabel 5. Metodes vir die ontleding van RGK in ’n opvoedkundige konteks

Outeur(s) en datum van publikasie	**Onderwerp**	**Kritiese besinning oor NVivo** **Ja ✓** **Nee ✗**	**Voordele van NVivo** **Ja ✓** **Nee ✗**
Abedin, Daneshgar en D’Ambra (2014)	Nagraadse studente se sosiale gedrag in asinchrone RGK	✗	✗
Clark, Couldry, MacDonald & Stephansen (2014)	Kollege studente se gesprekke via digitale platforms	✗	✗
Fleischmann (2014)	Tersiêre studente se gebruik van Flickr en Skype	✗	✗
Geng en Disney (2014)	Onderwysers se gebruik en kennis van SMS’e	✗	✗
Hewege en Perera (2013)	Die rol en implikasies van ’n wiki-gebaseerde pedagogie	✗	✗
Hillen (2014)	Die gebruik van gemeenskap-like besprekingsplatforms in digitale leeromgewings	✗	✗
Howard, Curwen, Howard & Colon-Muniz (2014)	Hoërskoolstudente se houdings teenoor ’n aanlyn sosiale netwerk platform	✗	✗
Junior, Gomes en Souza (2014)	Die gebruik van ’n sosiale netwerk in die onderrig van ’n vak in rekenaarwetenskap	✗	Voordele sluit in die fasilitering van die proses van datakategorisering en die ontwikkeling van hiërargiese bome.
Sabanci en Urhan (2014)	Hoërskoolstudente se gebruik van en standpunte oor sosiale media vir leerdoeleindes	✗	✗
Said, Forret en Eames (2014)	Die beperkinge van aanlyn medewerkende leerprosesse in ’n hoër onderwys instelling	✗	✗

Outeur(s) en datum van publikasie	**Onderwerp**	**Kritiese besinning oor NVivo** **Ja ✓** **Nee ✗**	**Voordele van NVivo** **Ja ✓** **Nee ✗**
Szeto en Cheng (2014)	Studente se ervaring van sosiale teenwoordigheid in 'n gemengde sinchrone leeromgewing	✗	✗
Hungerford-Kresser, Wiggins en Amaro-Jimenez (2014)	Die gebruik van blogs deur onderwysers	✗	✗
Pimmer, Brysiewicz, Linxen, Walters, Chipps & Gröhbiel (2014)	Mobiele leer in verpleegonderwys in landelike Suid-Afrika	✗	✗
Stewart (2014)	Die aanlyn geletterdheid van adolessente leerders van Engels buite skoolverband	✗	✗
Bruneel, Wit, Verhoeven & Elen (2013)	Gebruik in die onderwys en kwessies van privaatheid met betrekking tot Facebook in 'n hoër onderwys instelling	✗	✗
Deng en Tavares (2013)	Studente se motivering rakende aanlyn gemeenskappe in teme van Moodle en Facebook	✗	✗
Donnelly en Boniface (2013)	Onderwysers se persepsies en gebruik van 'n wiki om professionele ontwikkeling te bevorder	✗	✗
Çankaya, Durak en Yünkül (2013)	Hoekom voorgraadse studente opvoedkundige sosiale netwerk-webwerwe gebruik	✗	✗
Chan, Chu, Lee, Chan & Leung (2013)	Die gebruik van blogs en Facebook om kennis in 'n hoër onderwys instelling te bestuur	✗	✗
Menard-Warwick, Heredia-Herrera en Palmer (2013)	Die gebruik van internetgesprekke onder leerders van Engels as 'n vreemde taal	✗	✗
Mwalongo (2013)	Die gehalte en studente se persepsies van portuurgroepterugvoer in asinchrone gespreksforums	✗	✗

Outeur(s) en datum van publikasie	**Onderwerp**	**Kritiese besinning oor NVivo** **Ja ✓** **Nee ✗**	**Voordele van NVivo** **Ja ✓** **Nee ✗**
Nathans en Revelle (2013)	Kulturele diversiteit en temas in onderwysstudente se aanlyn besprekings	✗	✗
O'Brien en Glowatz (2013)	Die gebruik van Facebook as 'n akademiese instrument in hoër onderwys	✗	✗
Schoenborn, Poverjuc, Campbell-Barr & Dalton (2013)	Tersiêre studente se gebruik van web 2.0-toepassings	✗	✗
Snelson (2013)	Leerders se gebruik van blogs oor skool via video	✗	✗
Brannan en Bleistein (2012)	Beginner-onderwysers se persepsies van sosiale ondersteuningsnetwerke	✗	✗
Chen en Chen (2012)	Die gebruik van Twitter vir assesseringsdoeleindes	✗	✗
Chu, Siu, Liang, Capio & Wu (2013)	Nagraadse studente se ervarings en persepsies van wiki-platforms vir die bevordering van samewerkende leer en kennisbestuur	✗	✗
Crook (2012)	Die gebruik van web 2.0-gereedskap in sekondêre onderwys	✗	✗
Hickson (2012)	Maatskaplike werkers se selfbesinning via 'n blog	Hickson (2012:37) wys daarop dat daar vrae bly oor die gebruik van RGKDOP in kwalitatiewe navorsing.	Hickson (2012:37) merk op dat NVivo die kodering en ontleding, asook die herwinning, van data vergemaklik.
Kinash, Brand en Mathew (2012)	Studente se persepsies van mobiele leerprosesse	✗	✗
Pae (2012)		✗	Identifisering van voordele sluit in konsekwentheid wanneer dit kom by kodering en die fasilitering van die identifisering van tematiese eenhede.

Outeur(s) en datum van publikasie	**Onderwerp**	**Kritiese besinning oor NVivo** **Ja ✓** **Nee ✗**	**Voordele van NVivo** **Ja ✓** **Nee ✗**
Van Cleemput (2012)	Hoërskoolleerlinge se gebruik van kommunikasietegnologie (soos e-pos en kitsboodskappe) om te kommunikeer oor skoolverwante kwessies	✗	✗
Donnelly en Gardner (2011)	Inhoudsontleding van asinchrone RGK	✗	✗
Gallego-Arrufat, Gutiérrez-Santiuste en Campaña-Jiménez (2013)	Inhoudsontleding van onderwysers se gebruik van en nadink oor rekenaargesteunde onderrig	✗	✗
Nguyen (2011)	Studente se ervaring van RGK in ’n taalklas	✗	✗
Waterston (2011)	Ontleding van aanlyn interprofessionele gevallestudiebesprekings	✗	✗
Williams en Lahman (2011)	Studentebetrokkenheid en kritiese denke in aanlyn besprekings	✗	Williams en Lahman (2011:149) merk op dat hierdie instrument hulle gehelp het om hul koderingspraktyke te verfyn en te sinchroniseer.
Choi en Kang (2010)	Ontleding van aanlyn samewerkingsgroepwerk gedurende asinchrone RGK	✗	Daar is erkenning van NVivo as ’n instrument waarmee navorsers data kan kodeer op ’n volhoubare wyse.
Sidu en Embi (2010)	Die tersiêre student se rol in asinchrone RGK	✗	✗

MacMillan en Koenig (2004) noem ’n aantal redes waarom RGKDOP selde krities geëvalueer word, en noem onder andere die fokus op hierdie instrumente se tegnologiese vermoëns eerder as op metodologiese kwessies. Dit is beduidend om deur bogenoemde studies te lees en te let op hoe gereeld daar stellings gemaak word dat

NVivo gebruik is om die kodering of ontleding te ondersteun, sonder 'n eksplisiete verduideliking van watter metodes gevolg is (Jones & Diment 2010:82; Humble 2012:123). Jones en Diment (2010) skryf dat die gebruik van RGKDOP in die siening van sommige navorsers geldige metodologie en ontledingsmetodes kan vervang.[33] Trouens, deur gebruik te maak van inhoud-analise ondersoek Jones en Diment (2010) 325 kwalitatiewe navorsers se besigheid- en bestuur-georiënteerde artikels, en kom tot die gevolgtrekking dat 21% van hierdie artikels nie eksplisiete beskrywings van die navorsingsmetodes bied nie.

Die gebrek aan kritiese besinning kan deels spruit uit die mite dat 'n RGKDOP soos NVivo ontledingstake vir navorsers uitvoer, 'n wanindruk wat een van die grootste gevare vir die gebruik van RGKDOP in ernstige navorsing inhou (Matheson 2005:122). 'n RGKDOP kan slegs navorsers in die uitvoering van hul ontledings help; soos Bazeley en Jackson (2013:3) skryf kan rekenaarprogrammatuur nie slordige werk in goeie interpretasies omskep nie, en kan dit nie vergoed vir 'n beperkte begripsvermoë vanaf die navorser nie. Die onus lê op navorsers om hul data met integriteit te interpreteer, konseptualiseer en teoretiseer (Paulus, Lester & Britt 2013:640-641).[34]

4.3.2 Vrae van 'n *metodologiese aard*

Aansluitend by die vorige onderafdeling is daar soms vrae van 'n metodologiese aard, veral rakende die geldigheid van die gebruik van gekombineerde metodes (kwalitatief én kwantitatief). NVivo stel navorsers in staat daartoe om hul bevindinge in- en uit te voer na en van statistiese pakkette (Marshall & Friedman 2012:339), en bemiddel sodoende 'n gemengde-metodes benadering (Séror 2005:324). Navorsers wat gekant is teen die kombinasie van kwalitatiewe en kwantitatiewe metodes staan krities teenoor diegene wat RGKDOP gebruik, en hierdie metodes integreer volgens hulle op 'n arbitrêre wyse. MacMillan & Koenig noem dat sulke beskrywings effektiewelik kwantitatiewe en kwalitatiewe metodes as twee soortgelyke kategorieë verpak – een wat data in statistieke verander en die ander wat data in beskrywende kodes omskep. Hiervolgens word die metodes behandel as 'maar' twee navorsingsmetodes, een kwalitatief en die ander kwantitatief, en word die grense verder vervaag deur daarop te dui dat hierdie metodes versoenbaar genoeg is om saam gemeng te word. Hierdie teoretiese vaagheid,

33 Sien ook Bringer, Johnston & Brackenridge (2004:49), Leech (2010:267) en Clare (2012:4).

34 Natuurlik kan hierdie navorser-gerigte benadering beteken dat vooroordele aan die kant van die navorser kan voorkom.

waarin kwalitatiewe en kwantitatiewe metodes gemeng word en op 'n ad hoc basis gebruik word, stel RGKDOP in staat om binne die kategorie van kwalitatiewe ontleding geplaas te word, terwyl die deugde van 'n soort kwali-kwanti ontleding geloof word (MacMillan & Koenig 2004:182).

In teenstelling met navorsers wat sulke kritiese standpunte huldig, is daar 'n groot hoeveelheid navorsers wat wel gemengde metodes toepas (Friedman 2003; Paulus & Phipps 2008; Brokensha 2012). Marshall en Friedman (2012:34) wys daarop dat RGKDOP kwalitatiewe data in wese op 'n kwantitatiewe manier ontleed, met behulp van wiskundige algoritmes klassifiseer, tel, en andersins vergelykings maak, selfs in die afwesigheid van meting. Dit help om die kloof tussen kwantitatiewe en kwalitatiewe navorsing te oorbrug en daardeur 'n belangrike middeweg te vind. Die belangrike ding om in gedagte te hou, is dat indien navorsers 'n gemengde-metodes benadering tot ontleding onderneem, hulle versigtig moet wees en uitdruklik verwoord hoe dit inpas by die metodologiese perspektief wat hulle aanneem (Johnston 2006:384).

4.3.3 NVivo se beperkte toepassing binne gegronde teorie

Macmillan en Koening (2004:182) argumenteer verder dat selfs wanneer navorsers wat RGKDOP gebruik teorie erken, hulle geneig is om die geskiktheid van RGKDOP te evalueer deur die studie binne gegronde teorie te situeer. Daar bestaan 'n verkeerde oortuiging dat die gebruik van NVivo outomaties navorsers na gegronde teorie lei en dat die gereedskap ontwerp is om hierdie benadering te volg (Ozkan 2004:590; MacMillan & Koenig 2004:184; Clare 2012:4). NVivo is egter nie ontwerp met een spesifieke metode in gedagte nie. Moontlik is die verkeerde indruk dat NVivo ontwerp is om 'n ontleding deur middel van gegronde teorie te fasiliteer deels te danke aan die feit dat die term 'in vivo kodering' uit gegronde teorie afkomstig is en verwys na kategorieë en 'n benoeming deur middel van woorde wat deur mense self gebruik word (Richards 2009:104; Byrne & Callaghan 2014:199).

Navorsers moet ander metodes as slegs die gebruik van gegronde teorie verken, soos inhoud-analise, raamwerk-analise, narratiewe analise of fenomenografiese analise, om maar 'n paar te noem, en al hierdie benaderings en nog vele meer kan met NVivo uitgevoer word. Sou 'n mens byvoorbeeld 'vervreemding' bestudeer soos deur Seeman (1959) bespreek, sou 'n mens gevalle waar 'magteloosheid', 'betekenisloosheid', 'normloosheid', 'sosiale isolasie' en 'self-vervreemding' in 'n verskeidenheid tekste (onderhoude, persoonlike narratiewe, ensovoorts) voorkom binne NVivo kon kodeer. Nog 'n belangrike beginsel behels dat navorsers seker maak dat hulle geskikte teoretiese

perspektiewe vir 'n gegewe studie kies (Leech 2010:267), en vermy om perspektiewe te kies uitsluitlik op grond van wat hulle glo NVivo hulle kan bied (Bringer, Johnston & Brackenridge 2004:249). NVivo beperk nie die navorser nie, hy doen dit self.

4.3.4 Die data-ontledingsafstand

'n Beduidende kritiek teenoor RGKDOP soos NVivo is dat navorsers na bewering die risiko loop om van 'n meganiese, outomatiese ontleding van data gebruik te maak (Sinkovics & Alfoldi 2012:9), en ongelukkig blyk dit dat sommige navorsers wel met behulp van moderne RGKDOP teks outomaties kodeer vir 'n vinnige telling van reëlmaat (Bringer, Johnston & Brackenridge 2004:248). Die vrees bestaan dan dat die kwalitatiewe navorser homself nie meer verdiep in sy studie nie en ook soos 'n kwantitatiewe navorser 'n ontledingsafstand handhaaf.

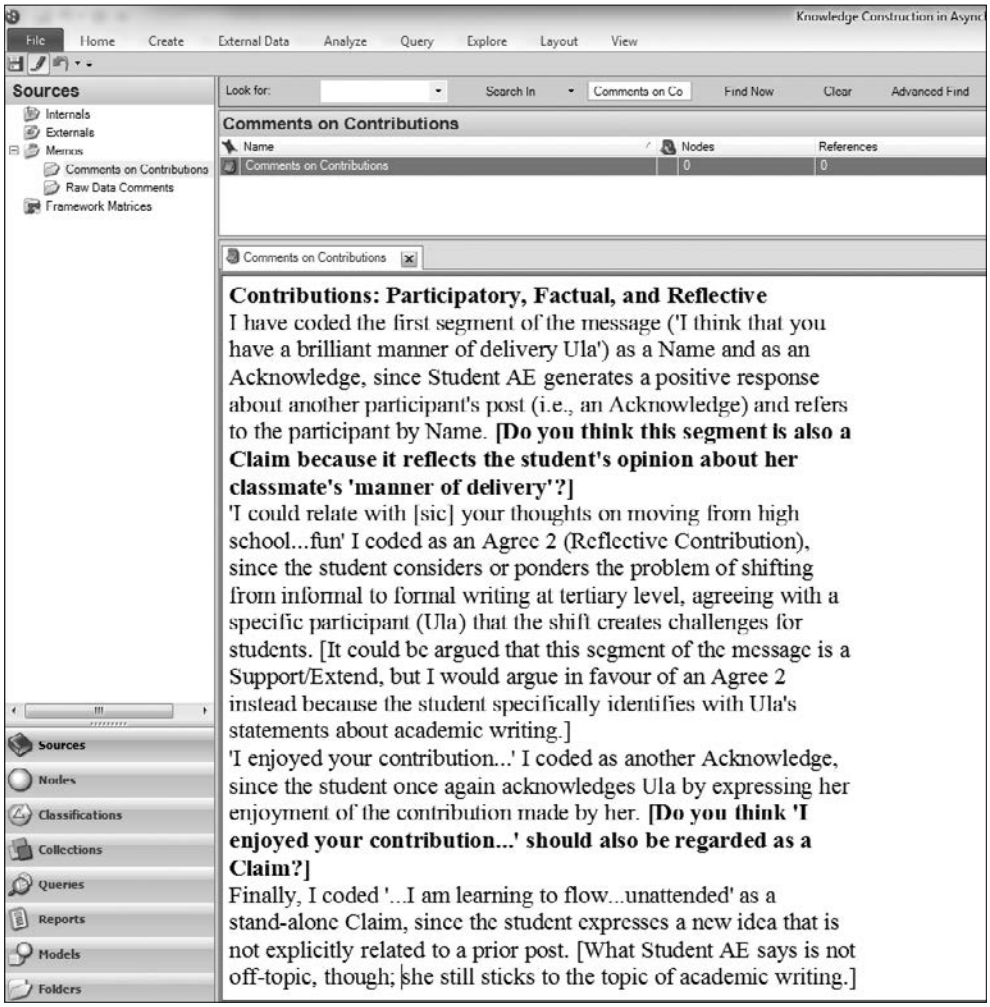
Die vrese van hierdie navorsers word verwoord deur Lichtman (2013:341), wat waarneem dat sy met haar eerste gebruik van rekenaarprogrammatuur bekommerd was dat sy besig was om in te koop in 'n paradigma wat waarde aan getalle, tafels en presisie heg, en dat kwalitatiewe rekenaarprogrammatuur dalk té gestruktureerd is. Hierdie kommer is ongegrond, aangesien 'n kwalitatiewe data-ontledingsinstrument soos NVivo nie, soos vroeër opgemerk, die werklike ontleding vir die navorser behartig nie – dit is steeds die navorser wat die ontleding uitvoer en interpreteer (Humble 2012:125). NVivo en ander RGKDOP bied juis 'n manier vir die geesteswetenskaplike navorser om in gesprek te tree met grootdata sonder om 'n paradigmaskuif na die kwantitatiewe te onderneem.

Om die menslike interpretasie van data, en interne geldigheid (betroubaarheid) van 'n gegewe studie, te verseker kan ontleders wat NVivo gebruik 'n aantal metodes aanwend, onder andere deeglike beskrywing en triangulering. Volgens Tracy (2010:843) behels deeglike beskrywing – 'n kwalitatiewe navorsingskonsep geskep deur die antropoloog Clifford Geertz (1973:26-27) – dat navorsers gedetailleerde beskrywings van bepaalde verskynsels soos sosiale gebeure of aksies verskaf sodat ander navorsers hul eie gevolgtrekkings kan bereik en vasstel of dit geldig is om 'n bepaalde hipotese op 'n gegewe navorsingsomgewing van toepassing te maak. NVivo het 'n memoranduminstrument waarmee navorsers digte, noukeurige beskrywings van hul data kan aanteken. Twee belangrike voordele hiervan is dat navorsers hulself kan verdiep in hul data, en indien hulle as deel van 'n span werk kan hulle ook insig in mekaar se interpretasies verkry. Figuur 10 illustreer die gebruik van NVivo se memorandum deur een van die skrywers van hierdie boek. Haar doel is onder andere om haar denke, interpretasies en gevolgtrekkings met haar mede-navorser te deel.

Betroubaarheid word versterk omdat die opstel van memorandums die ontleder in staat stel om sy of haar metodologiese prosesse deursigtig te maak (Ryan 2009:158). Daarbenewens laat memorandums navorsers toe om 'n ontledingsouditspoor na te laat, wat veral noodsaaklik is in kwalitatiewe navorsingsopsette wat gekenmerk word deur subjektiwiteit en navorsers se vooroordele (Petty, Thomson & Stew 2012:381).

Die tweede metode waardeur betroubaarheid versterk kan word, triangulering, word bereik deur middel van, onder andere, verskeie metodes, teoretiese raamwerke, en databronne, ongeag of kwalitatiewe data-ontledingsprogrammatuur gebruik word of nie. Datatriangulering is maklik om te bewerkstellig met behulp van NVivo, aangesien onderhoude, waarnemingsnotas, en argiefstukke in NVivo ingesluit kan word, terwyl navorsertriangulering bereik kan word deur NVivo se memoranduminstrument, aangesien navorsers kan vergelyk en/of hul kodering kan verfyn, verskille identifiseer, en konsensus onderhandel.

In die memorandum in bogenoemde figuur is dit duidelik dat die navorser deur middel van self-refleksie nie net haar kodering van 'n asinchrone boodskap in detail met haar mede-navorser bespreek nie, maar ook deur middel van vrae aandui dat sy 'n paar onsekerhede oor haar kodering ervaar. Sulke deeglike beskrywings via NVivo kan nie alleenlik navorsers se lyste van kodes oorskadu nie (Polit & Beck 2010:1456), maar moedig ook deursigtigheid aan deur die meedeel van die studie se uitdagings en onverwagse kinkels en draaie (Tracy 2010:842; Brokensha & Greyling 2014).



Figuur 10. Deeglike beskrywings met behulp van NVivo

4.4 Gevolgtrekking

NVivo het baie te bied aan navorsers wat hul kwalitatiewe data-ontleding wil verbeter. NVivo gaan duidelik verder as die blote kodering en herwinning van data, omdat dit tot baie ander strategieë lei, insluitend die optekening van memorandums, self-refleksie, en die verfyning van koderingskategorieë (Beekhuyzen, Nielsen & Heller 2010:4). Dit beteken nie dat navorsers voor die voet die lof van NVivo moet besing nie; 'n strik waarin beginner-RGKDOP-navorsers hulself soms bevind. Bong (2007:259) stel dit

bondig wanneer sy sê dat sy as 'n 'digitale immigrant' aanvanklik mislei is deur die skone nuutheid van die gebruik van RGKDOP, en dat sy daarvan gehou het dat dit ooreengestem het met die oorspronklikheid van haar navorsing. Aangesien sagteware soos NVivo nie werklik die ontleding vir die navorser voltrek nie, kan dit in wese nie beskou word as 'n metode van interpretasie nie. Eerder as om NVivo as 'n instrument te sien wat metodes dryf – soos tegnologiese deterministe dit sien – is dit meer nuttig om NVivo te benader as 'n instrument met funksies wat gemik is op die ondersteuning van 'n spesifieke metode (Gibbs 2014:277-278).

NVivo verteenwoordig 'n voorbeeld van wat die kwalitatiewe navorser in 'n era van grootdata kan vermag. Nie alleen kan dié programmatuur groot datastelle binne 'n enkele projek akkommodeer nie, maar is NVivo ook spesifiek ontwerp om om te gaan met die verskeidenheid van bronne wat kenmerkend van die grootdatawêreld is. Met behulp van dié programmatuur kan die kwalitatiewe navorser steeds akkuraat met sy data omgaan, maar op 'n baie groter skaal, en kan hy die ryk verskeidenheid van bronne wat kenmerkend van 'n kwalitatiewe navorsingsprojek is akkommodeer. Daar bestaan wel kritiek teen RGKDOP oor die algemeen, maar soos in dié hoofstuk aangedui is, kan NVivo baie van die kritiek aanspreek.

Benewens die gemengde kwalitatiewe/kwantitatiewe benaderings wat in dié hoofstuk genoem is, onderneem Long, Cunningham, Wiley, Carswell en Braithwaite (2013) 'n interessante studie in hul ontleding van semi-gestruktureerde onderhoude met NVivo, maar gebruik UCINET om netwerkberekenings te doen, en NetDraw om netwerke te visualiseer. Netwerkontleding is die onderwerp van die volgende hoofstuk.

Hoofstuk 5

Netwerkontleding

5.1 Inleiding

Netwerkontleding is een belangrike wyse waarop die verhoudinge in grootdatastelle visueel voorgestel kan word om ontledings te fasiliteer en bevindinge te kommunikeer (Fox & Hendler 2011:707). Die interdissiplinêre toepassingsmoontlikhede van netwerkontleding, tesame met die groeiende gewildheid van hierdie benadering binne die wetenskap (die joernaal *Complex Networks* is byvoorbeeld eers in 2013 gestig), die belangrikheid van netwerkontleding in grootdatabenaderings, en netwerke se konsepsuele oorvleueling met die teorieë van kompleksiteit en sisteme, beteken dat hierdie ook 'n noemenswaardige metode is waarvan navorsers kennis moet neem. Hierdie hoofstuk fokus op die visuele voorstelling en verkenning van netwerke, aangesien die onderliggende berekenings reeds elders breedvoerig behandel is (Senekal 2014b).

Netwerkteorie se oorsprong kan sover nagespoor word as Leonard Euler se bekende Königsberg-probleem van 1736, maar het eers onlangs werklik byval begin vind as 'n wetenskaplike benadering. Borgatti, Mehra, Brass en Labianca (2009:892) beskryf dit as brandende kwessie, terwyl Lima (2011:221) skryf dat die netwerkteorie in die 'kern' van die wetenskaplike revolusie staan. Veral vier faktore het bygedra daartoe dat die netwerkteorie vandag so gewild is:

- Fisici se betrokkenheid by dié veld sedert die seminale studies van Watts en Strogatz (1998) en Barabási en Albert (1999),
- die beskikbaarheid van groot en betroubare digitale datastelle,
- beter en sterker rekenaars, en
- die sogenaamde 'globale oorlog teen terreur'.

Laasgenoemde het 'n belangrike finansiële inspuiting in die veld tot gevolg gehad, aangesien SNA ook gebruik word in die ontleding van terrorisnetwerke: Die *US Army / Marine Corps counterinsurgency field manual* (2006) wy byvoorbeeld 'n afdeling aan SNA wanneer oor militêre intelligensie geskryf word. Voorbeelde waar SNA toegepas is in die globale oorlog teen terreur sluit in die soeke na skakels tussen terroriste wat betrokke was by die aanvalle op die Wêreldhandelsentrum in New York

op 11 September 2001, die ontleding van die bomaanvalle in Madrid in Maart 2004, en die opsporing van Saddam Hussain (sien onderskeidelik (Krebs 2002; Department of the Army and Department of the Navy 2006:B45 e.v.; Ressler 2006:3-4).[35] Alhoewel die oorsprong van SNA dus ver in die verlede lê, het militêre intelligensie oor die afgelope dekade 'n groot rol gespeel in die benutting en ontwikkeling van hierdie veld, met die praktiese toepassing daarvan wat geïllustreer het dat SNA 'n bruikbare benaderingswyse verskaf wat ontleders in staat stel om sleutelfigure in groot datastelle te help identifiseer (Department of the Army and Department of the Navy 2006:B40). As sodanig is die netwerkteorie uniek in die opsig dat dit nie alleen multidissiplinêr is nie, maar ook in die praktyk sowel as in verskeie akademiese velde aangewend word.

Volgens Scott (1996:211) is die groei van SNA direk verwant aan die ontwikkeling van rekenaarprogrammatuur. Binne die sosiologie wys Boissevain (1979:392) reeds daarop dat SNA die geleentheid geskep het om data rekenaarmatig te ontleed, en Tichy, Tushman en Fombrun (1979:513) noem die programme DIP, SocPac, SOCK, COMPLT, BLOCKER en CONCOR, terwyl Haythomthwaite (1996:331) na GRADAP, STRUCTURE, UCINET, NEGOPY en KRACKPLOT verwys.[36] Senekal (2012a) noem ook die gratis program, NetDraw, wat deur Steve Borgatti ontwikkel is. Die akademiese standaard is Pajek en UCINET, maar enige SNA-program kan 'n netwerkontleding behartig. Programmatuur wat veral binne die veld van militêre intelligensie aangewend word sluit in Sentinel Visualizer, i2 Analyst Notebook, Starlight VIS en Palantir. In die ontledings wat volg, word Gephi gebruik, wat ontwikkel is deur Bastian, Heymann, en Jacomy (2009) en onder andere bespreek word in Heymann en Le Grand (2013) en Cherven (2013).

SNA fokus op rolle en posisies, hoe 'n netwerk gestruktureer is, hoe invloed en mag versprei, en hoe hulpbronne benut word. 'n Ontleder kan met behulp van SNA vinnig 'n oorsigtelike blik van skakels kry en sleutelfigure uitlig, wat beteken dat nuwe insigte ontdek kan word deur die grafiese voorstelling van netwerke. Netwerkteorie deel ook die uitkyk van die sisteemteorie deurdat die fokus op skakels tussen entiteite binne 'n sisteem of netwerk val, eerder as op die entiteite self. Amaral en Ottino (2004:147) merk op dat die netwerkteorie een van die mees sigbare benaderingswyses geword het by die beskrywing, ontleding en begrip van komplekse sisteme.

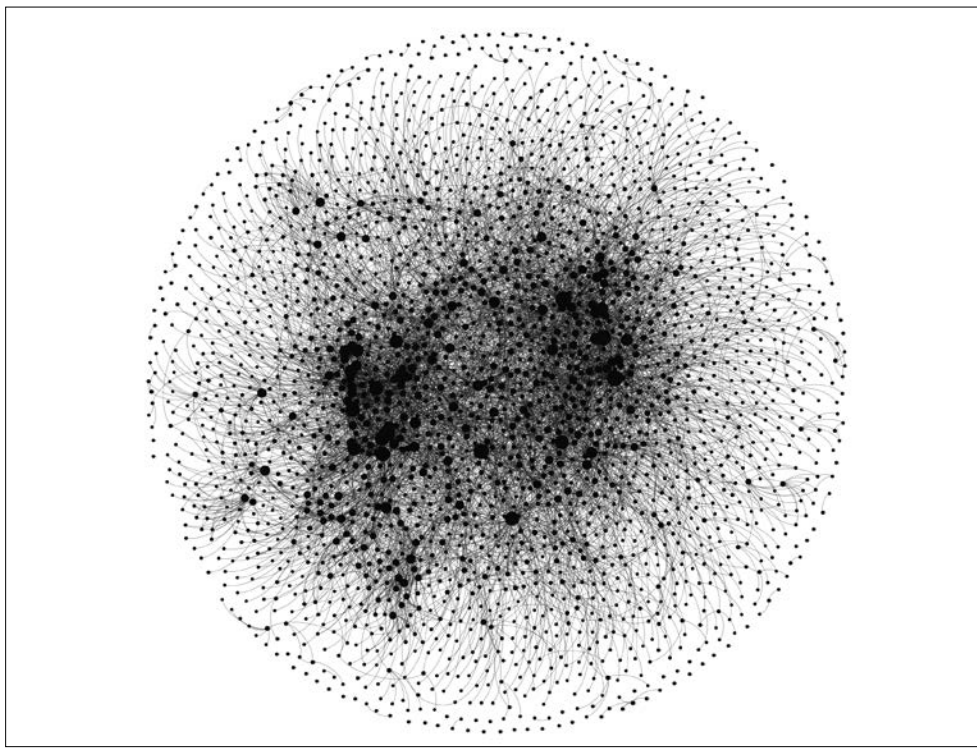
35 Sien ook Rodriguez (2005) en Senekal (2014c) oor die opkoms van SNA binne die raamwerk van die globale oorlog teen terreur.

36 Sien ook Scott (1996:212).

'n Netwerkgrafiek (ook 'n 'sosiogram' genoem waar sosiale netwerke ter sprake is) is 'n voorstelling van die aktiwiteite van entiteite (ook genoem 'akteurs' of 'nodusse') en van die skakels tussen hulle. Bykans enige sisteem kan as 'n netwerk voorgestel word, en Newman (2003, 2010) onderskei tussen vier soorte netwerke: biologiese netwerke, tegnologiese netwerke, inligtingsnetwerke, en sosiale netwerke.

5.2 Biologiese netwerke

Biologiese netwerke sluit in metaboliese prosesse, proteïeninteraksies, ekosisteme, senuweenetwerke, ensovoorts. So kan 'n voedingsnetwerk byvoorbeeld opgestel word, waar nodusse verwys na spesies en die skakels voedingspatrone tussen spesies verteenwoordig. Alhoewel die navorser binne die geesteswetenskappe nie met hierdie soort netwerke sal werk nie, is dit belangrik om ook na een van hierdie netwerke te kyk, aangesien bykans alle komplekse netwerke ooreenstemmende kenmerke vertoon. Die grafiek in Figuur 11 is 'n voorstelling van die interaksies tussen proteïne in die gis *Saccharomyces cerevisiae* (data verskaf deur Bu et al. (2003)).



Figuur 11. Die interaksies tussen proteïne in *Saccharomyces cerevisiae*

Let daarop dat daar meer nodusse en skakels in die sentrum van hierdie netwerk voorkom en minder op die periferie. Kraggebaseerde uitlegte,[37] wat aandui tot watter mate nodusse ander aantrek, word gereeld gebruik in die visualisering van netwerke (Fruchterman en Reingold se uitleg is hier aangewend). Die nodusse wat die mees sentrale rol in 'n netwerk speel, word in die sentrum geposisioneer, terwyl minder sentrale nodusse – wat ook gewoonlik minder skakels het – op die periferie geposisioneer word (Kobourov 2013:397). Dit beteken dat die netwerkteorie ook gebruik kan word om te bepaal of 'n nodus binne die sentrum of op die periferie van 'n netwerk funksioneer.

5.3 Tegnologiese netwerke

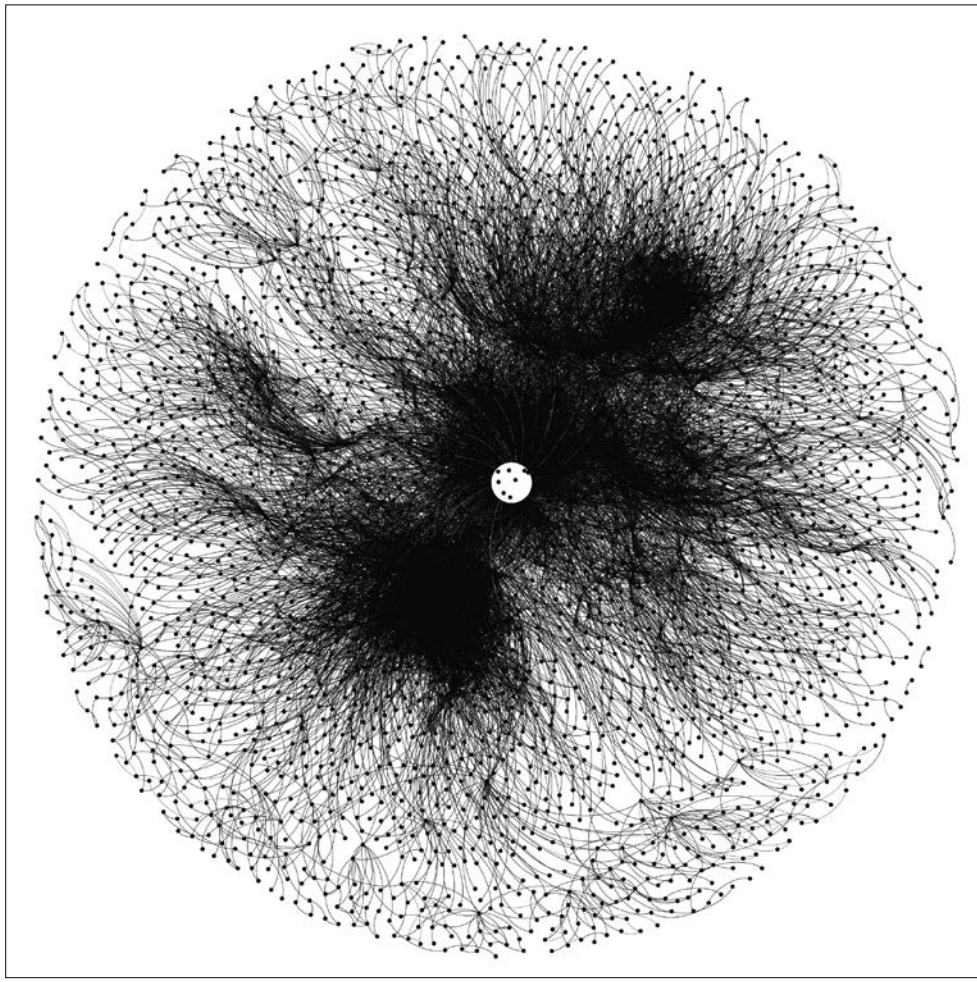
Tegnologiese netwerke sluit in kragvoorsieningsnetwerke, vervoernetwerke (byvoorbeeld die van internasionale vlugte), en die internet. Hoewel die wêreldwye web 'n inligtingsnetwerk is en die internet 'n tegnologiese netwerk, vertoon hulle 'n soortgelyke netwerkstruktuur omdat beide komplekse netwerke is. Dit sluit in die teenwoordigheid van nodusse met meer skakels as ander (die sterstrukture in hierdie netwerke); 'n gemiddelde kortpad tussen alle nodusse ('klein-wêreldsheid', soos deur Watts en Strogatz (1998) geïdentifiseer); 'n kragwetverspreiding van skakels (soos deur Barabási en Albert (1999) geïdentifiseer); en selektiewe skakelvorming van nodusse volgens graadkorrelasie of homofilie (soos veral deur Newman (2002) bestudeer). Die wêreldlugvaartnetwerk word in Figuur 12, as 'n voorbeeld van 'n tegnologiese netwerk[38] waarin 2 988 lughawens deur 15 643 vlugte verbind word, aangedui (Londen se Heathrow is interessantheidshalwe met wit aangedui).

Weereens kan gesien word dat sommige nodusse (in hierdie geval lughawens) meer skakels het as ander, soos aangedui deur hul grootte. Die ontleding van tegnologiese netwerke word onder andere gebruik in die bestudering van die verspreiding van siektes, aangesien moderne vervoernetwerke juis vinnige wêreldwye verspreiding bevorder. Ten opsigte van kragvoorsieningsnetwerke was van die belangrikste bevindings van sulke netwerkanalises dat komplekse netwerke veerkragtig is; met ander woorde as sommige skakels verwyder word (wanneer 'n kragsentrale byvoorbeeld deur 'n orkaan verwoes word) kan hierdie netwerke steeds met minimale ontwrigting funksioneer, maar dat hulle kwesbaar is vir aanvalle wat die belangrikste nodusse in die netwerk verwyder. Hierdie

37 Onder andere ontwikkel deur Eades (1984), Kamada en Kawai (1989), Fruchterman en Reingold (1991), en Hu (2011).

38 Data verkry vanaf http://openflights.org

insig het ook ’n belangrike invloed uitgeoefen op sosiale netwerke, waar die oogmerk van die toepassing van dié benadering op terroristenetwerke juis is om die belangrikste nodusse te identifiseer en te elimineer, en sodoende die netwerk te laat disintegreer.



Figuur 12. Die wêreldlugvaartnetwerk (Heathrow in Londen word interessantheidshalwe met wit aangedui)

5.4 Inligtingsnetwerke

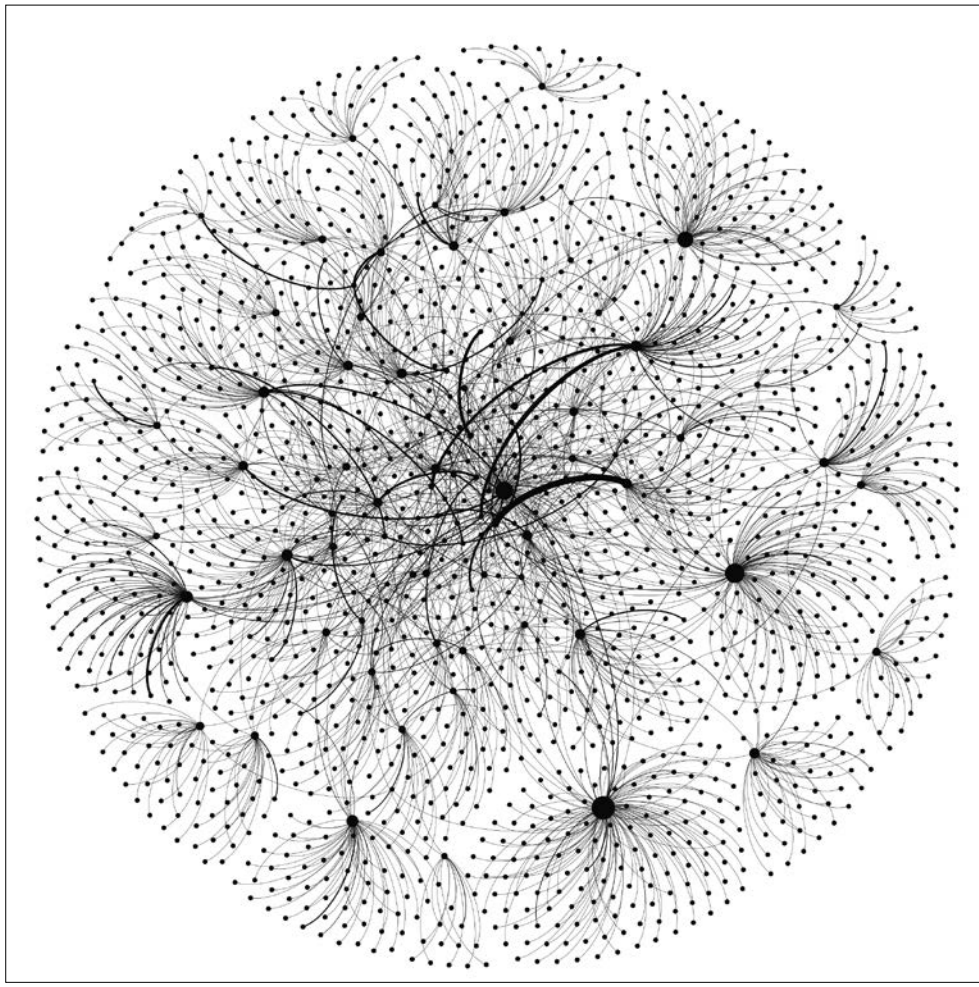
Die derde soort netwerk, inligtingsnetwerke, sluit die wêreldwye web en verwysingspatrone van akademiese artikels in. In die ontleding van die wêreldwye web verteenwoordig die nodusse webblaaie, en die skakels is dan die tussen webblaaie. So sou 'n mens byvoorbeeld die netwerk van webblaaie kon ondersoek waarbinne die diskoers rondom die Afrikaanse letterkunde plaasvind.

Verwysingsontleding is een van die 'klassieke' toepassings van die netwerkteorie (Newman 2003:176). Hierdie ontledings dui aan watter outeurs en akademiese joernale die meeste aangehaal word, watter outeurs die meeste bronne aanhaal, waar hulle oorvleuel met ander outeurs, en watter joernale en outeurs in die kern van 'n akademiese veld funksioneer. So kan byvoorbeeld bepaal word watter teoretici en akademiese joernale die grootste invloed uitoefen binne 'n dissipline, hetsy deur 'n berekening van die aantal kere wat 'n outeur aangehaal word of deur te sien of 'n outeur binne die sentrum of op die periferie van die netwerk geposisioneer word. Figuur 13 dui die verwysingsnetwerk van akademiese artikels binne die Afrikaanse letterkunde van 2011 tot 2012 aan.[39]

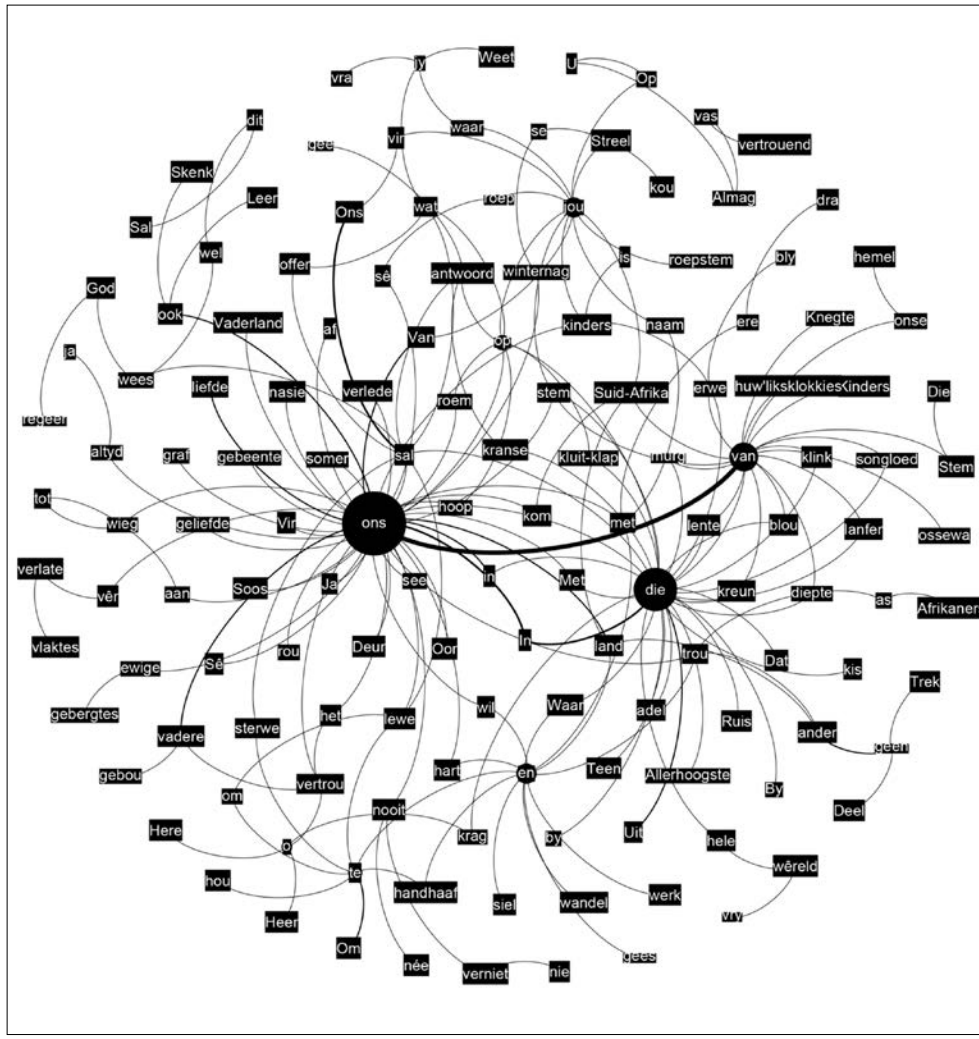
Soos in die geval van tegnologiese netwerke ontwikkel daar 'n sterstruktuur om sommige nodusse, wat daarop dui dat hierdie nodusse veral baie skakels met ander nodusse in die netwerk het. Om die netwerk verder te bestudeer kan dit vergroot of vereenvoudig word (deur die verwydering van sommige soorte nodusse; joernale in hierdie geval), en wiskundige berekenings kan gedoen word om die rol van individuele nodusse te bepaal.

Taal kan ook gesien word as 'n inligtingsnetwerk. In 'n netwerkontleding word dan veral ondersoek ingestel na die patrone van interaksies tussen woorde, hetsy op 'n semantiese (byvoorbeeld sinonieme of antonieme) of op 'n sintaktiese (byvoorbeeld watter woorde saam voorkom) vlak. Figuur 14 dui die leksikale netwerk in C.J. Langenhoven se "Die stem", waar die skakels tussen woorde aandui of hulle langs mekaar voorkom, aan.

39 Data verkry vanuit Senekal (2014d).



Figuur 13. Die verwysingsnetwerk van akademiese artikels binne die Afrikaanse letterkunde (2011-2012)



Figuur 14. Die leksikale netwerk in "Die stem"

Hieruit kan gesien word dat 'ons' die woord is wat op 'n sintaktiese vlak 'n sleutelrol speel deur die grootste aantal woorde saam te bind. Aangesien 'die' gewoonlik hierdie rol vervul in die meeste ander tekste (en in Engels word hierdie rol ook vervul deur die bepaalde lidwoord), is 'ons' se sentrale posisie hier van besondere belang – op 'n sintaktiese vlak beeld "Die stem" die idee van samehorigheid uit, soos ook in die leuse van die ou Suid-Afrika gevind word ("Ex unitate vires"). Die dikker lyn tussen 'van' en 'ons' dui om die beurt daarop dat hierdie twee woorde gereeld langs mekaar

voorkom. Die leksikale netwerk in "Die stem" is natuurlik slegs 'n aanduiding van hoe woorde in hierdie teks skakel, en 'n mens sou groter en meer tekste moes ontleed (en verkieslik ook werklike, eerder as literêre, taalgebruik) om vas te stel hoe woorde in Afrikaans saamhang.

5.5 Sosiale netwerke

Sosiale netwerke is die tradisionele domein van SNA en sluit samewerking tussen wetenskaplikes, vriendskapsnetwerke, organisatoriese strukture, netwerke van maatskappydirekteure, en familiebande in. Die netwerk van wetenskaplikes wat saam met Paul Erdös gepubliseer het, is byvoorbeeld al op hierdie manier bestudeer.[40] Erdös was een van die mees produktiewe navorsers van die 20ste eeu en het meer as 1400 artikels in sy leeftyd gepubliseer. 'n Groot aantal van hierdie publikasies was saam met medeouteurs, onder wie baie ook saam met ander wetenskaplikes gepubliseer het. Op hierdie manier het 'n netwerk van samewerking ontstaan.

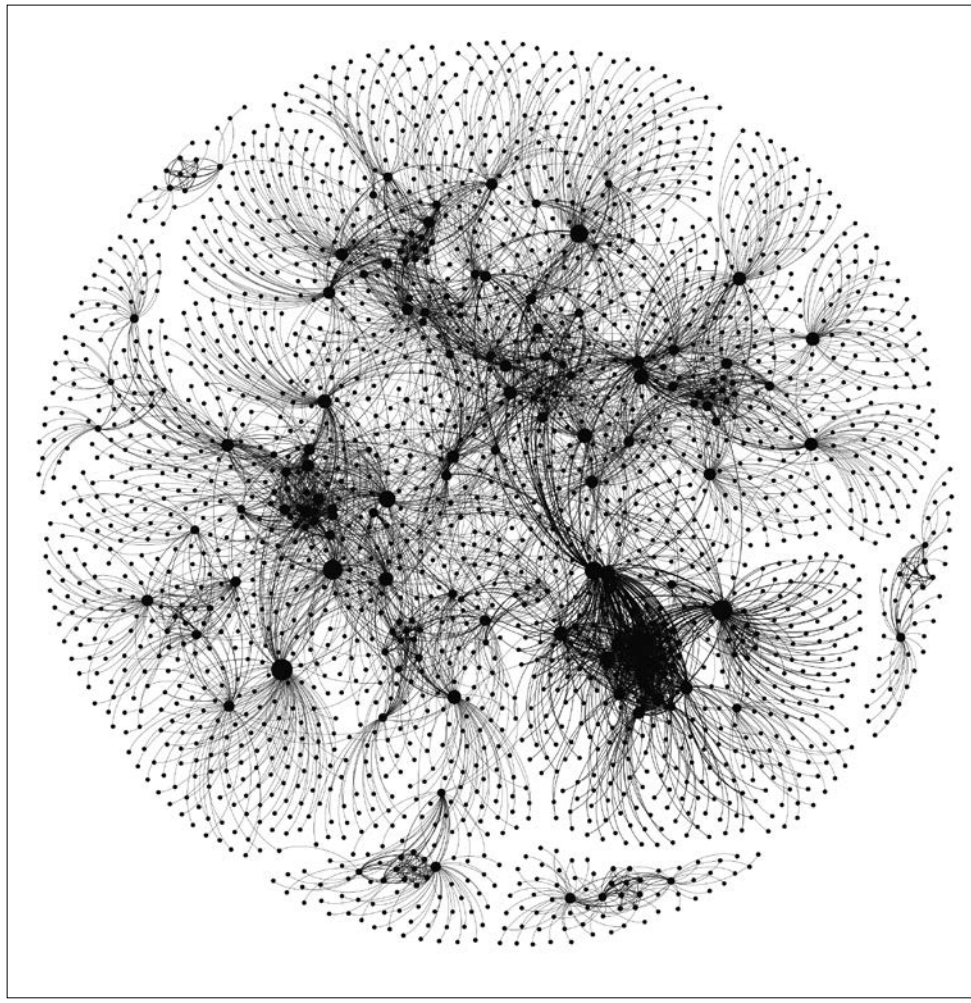
Verskeie sosiale netwerke kan só bestudeer word. As voorbeeld word die netwerk van Suid-Afrikaanse direkteure in die plaaslike bankwese[41] in Figuur 15 aangedui.

Die navorser kan die skakels tussen rolspelers verder ondersoek, die netwerk vergroot, of wiskundige berekenings doen om die rol van individuele nodusse te bepaal.

Die Afrikaanse literêre sisteem word in Senekal (2013, 2014) ook as 'n sosiale netwerk bespreek waar mense binne dié veld skryf en oor geskryf word. Die grafiek in Figuur 16 is 'n voorstelling van die Afrikaanse literêre sisteem vanaf 1900 tot 1978, 'n netwerk wat bestaan uit 3 641 entiteite met 14 507 verbindings, met data verkry uit Senekal en Van Aswegen (1980, 1981) en Senekal en Engelbrecht (1984).

40 Sien Buchanan (2003:34-35), Watts (2004:93) en Strogatz (2004:246-247).

41 Data verkry vanuit Senekal en K. Stemmet (2014).



Figuur 15. Die Suid-Afrikaanse bankdirekteurnetwerk

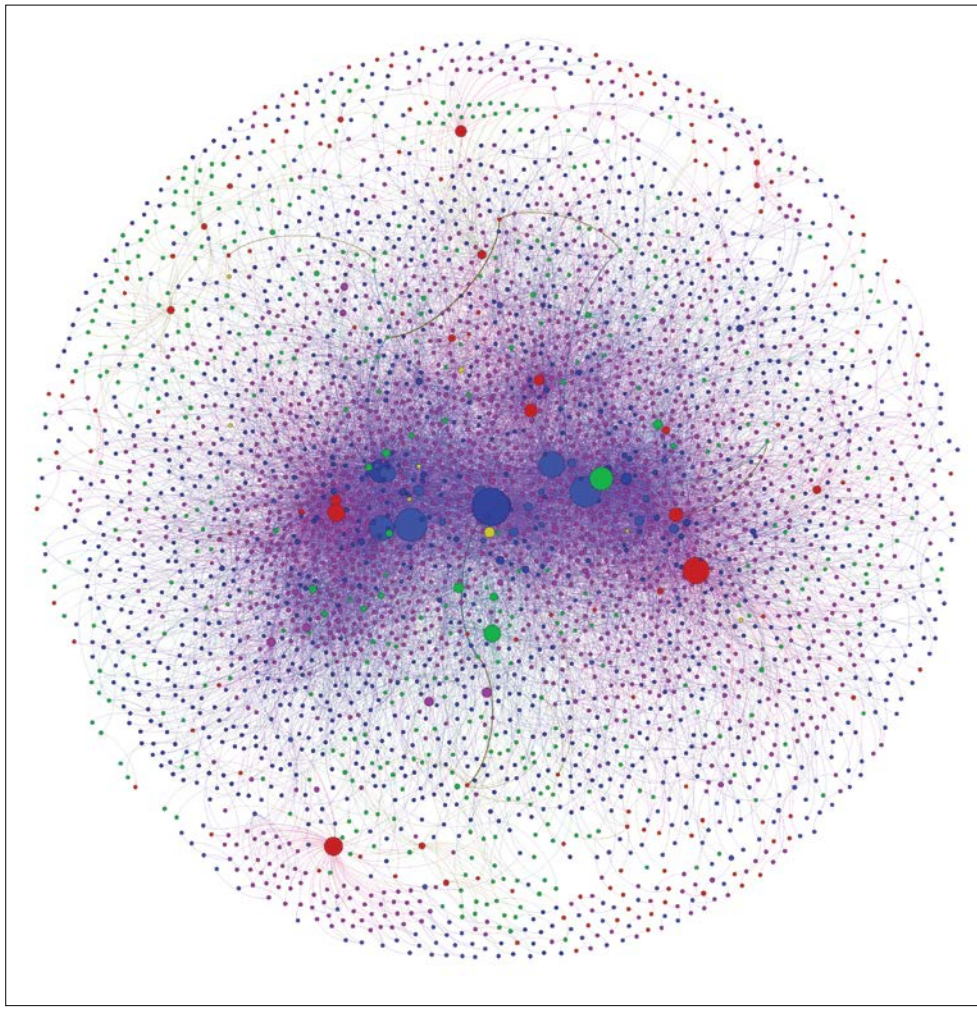
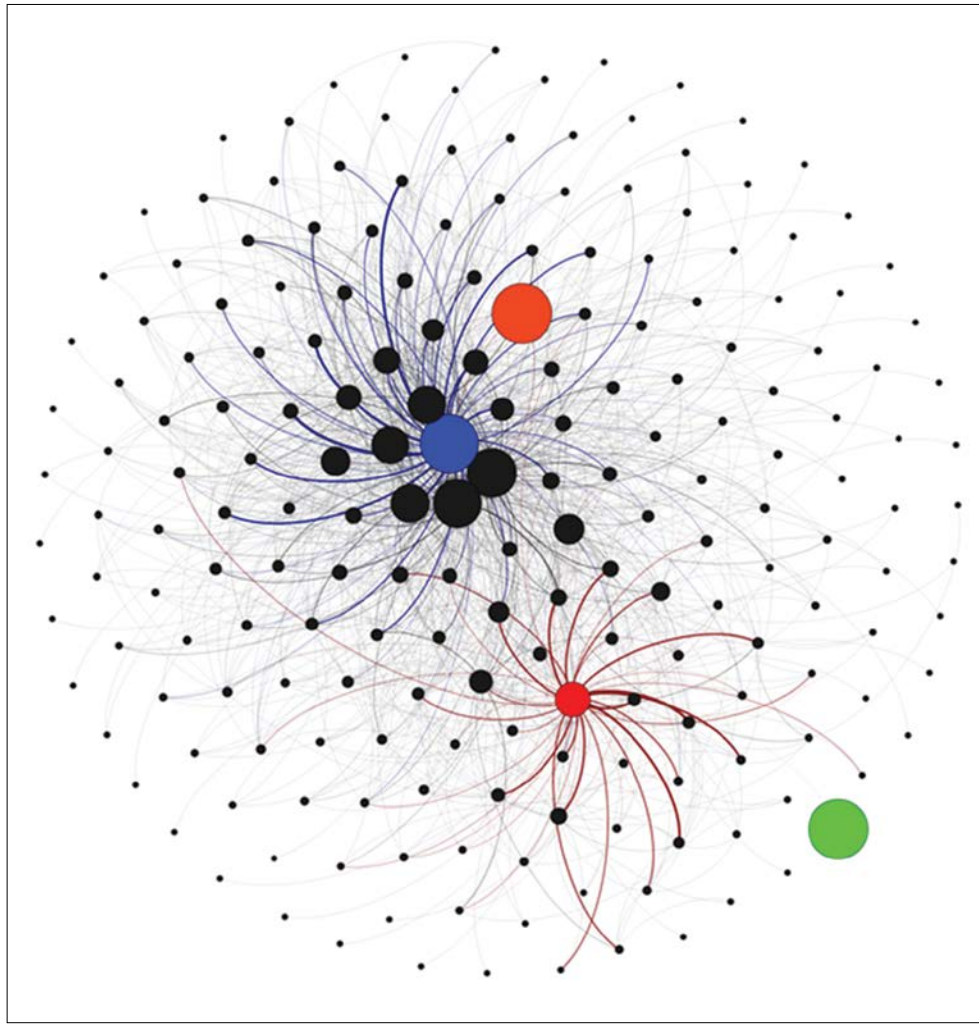


Figure 16. Die Afrikaanse literêre sisteem (1900-1978)

Omdat verskeie rolspelerkategorieë hierby betrokke is, is die verskillende nodusse ook gekleur: werke in pienk, mense in blou, uitgewerye in rooi, publikasieplatforms (koerante, joernale en tydskrifte) in groen, en pryse in geel. Deur verskillende soorte rolspelers met behulp van verskillende kleure aan te dui kan die navorser makliker met die netwerk omgaan. Hier kan byvoorbeeld gesien word dat daar uitgewerye binne die sentrum van die sisteem funksioneer, terwyl ander op die periferie besig is. By nadere ondersoek blyk dit dat DALRO en selfpublikasies op die periferie geposisioneer is, terwyl Naspers, Tafelberg, Human & Rousseau en ander binne die sentrum aangetref

word. Só kan binne 'n enkele oogopslag gesien word watter rolspelers die belangrikste posisies beklee vir die funksionering van dié sisteem.

Ekonomiese netwerke word ook gewoonlik as sosiale netwerke geag, en kan ook lig werp op die geskiedenis. Neem byvoorbeeld die internasionale wapenhandelnetwerk vanaf 1948 tot 1989, met data verskry vanaf SIPRI (Stockholm Institute for Peace Research). In Figuur 17 is die VSA blou gekleur, die Sowjetunie rooi, Suid-Afrika oranje en die ANC groen (Suid-Afrika en die ANC is vergroot ter wille van duidelikheid).



Figuur 17. Die internasionale wapenhandelnetwerk (1948-1989)

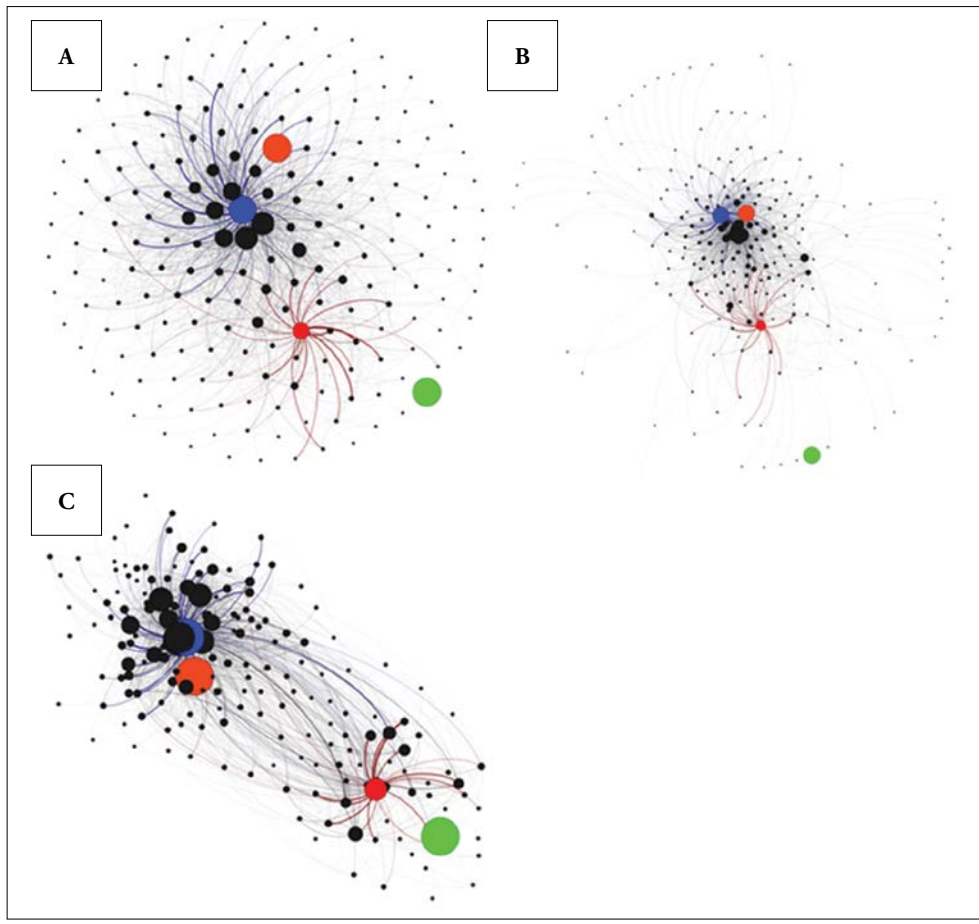
Hier kan gesien word dat veral twee groeperings vorm: rondom die VSA, en rondom die Sowjetunie. Die wapenhandelnetwerk stel dus ’n duidelike Koueoorlogse digotomie voor, met Suid-Afrika in die VSA-kamp en die ANC in die kamp van die Sowjetunie. Dié uitleg is gedoen met behulp van Fruchterman en Reingold (1991) se kraggebaseerde uitlegalgoritme, en ’n mens sou groeperings duideliker kon voorstel met behulp van OpenOrd (Martin, Brown, Klavans & Boyack 2011), soos in die volgende afdeling bespreek word.

5.6 Uitlegalgoritmes

Verskeie uitlegalgoritmes is reeds binne die netwerkteorie ontwikkel om komplekse sisteme op ’n sinvolle wyse te kan voorstel sodat patrone van interaksies ontdek kan word. Die gewildste hiervan vir akademiese ondersoek is kraggebaseerde uitlegalgoritmes, wat die sisteem of netwerk as ’n fisiese sisteem benader en skakels as kragte hanteer wat entiteite aantrek/afstoot totdat ’n toestand van ewewig bereik word (Hu 2011:40; Van Steen 2010:47-49; Merico, Gfeller & Bader 2009:922; Gaertler & Wagner 2007:117; Suderman & Hallett 2007:2654). Christakis en Fowler (2007) het byvoorbeeld Kamada en Kawai (1989) se uitlegalgoritme benut in hul studie van die verspreiding van vetsug, soos ook Vicarelli, De Benedictis, Nenci, Santoni en Tajoli (2013) in hul bestudering van wêreldhandelsnetwerke. Die voordeel van dié algoritmes is dat gesien kan word watter entiteite sentraal binne ’n netwerk funksioneer. Kyk byvoorbeeld weer na die netwerk wat die internasionale wapenhandelnetwerk voorstel: die VSA en die Sowjetunie was natuurlik van die belangrikste rolspelers in dié industrie gedurende die Koueoorlog, en hul sentrale posisies kan deur die grafiese voorstelling van die netwerk uitgelig word. Hierteenoor was die ANC slegs aan die ontvangkant van wapentransaksies, met geen industrie van sy eie nie, wat visueel uitgebeeld word deur ’n periferale posisie.

Grootdata stel nuwe eise aan beide navorsers en rekenaarprogrammatuur, veral aangesien ontledingsmetodes wat op ’n klein skaal toegepas kan word dikwels nie op ’n groot skaal werk nie (Fan, Han & Liu 2014:13-16). ’n Voorbeeld is Fruchterman en Reingold se kraggebaseerde uitlegalgoritme soos hierbo gebruik is. Alhoewel dit baie bruikbaar is vir die visualisering van netwerke met ’n paar honderd entiteite, kan dit nie groot netwerke met tienduisende of miljoene entiteite en hul skakels hanteer nie (Hu 2011:38; Martin et al. 2011:2). Verder kan dit ook nie groeperings duidelik uitlig nie, wat beteken dat aanwending beperk is tot die identifisering van sentrale en periferale rolspelers. Om hierdie redes is verbeterde uitlegalgoritmes wat spesifiek met groter datastelle kan omgaan deur onder andere Hu (2011) en Martin et al. (2011) voorgestel.

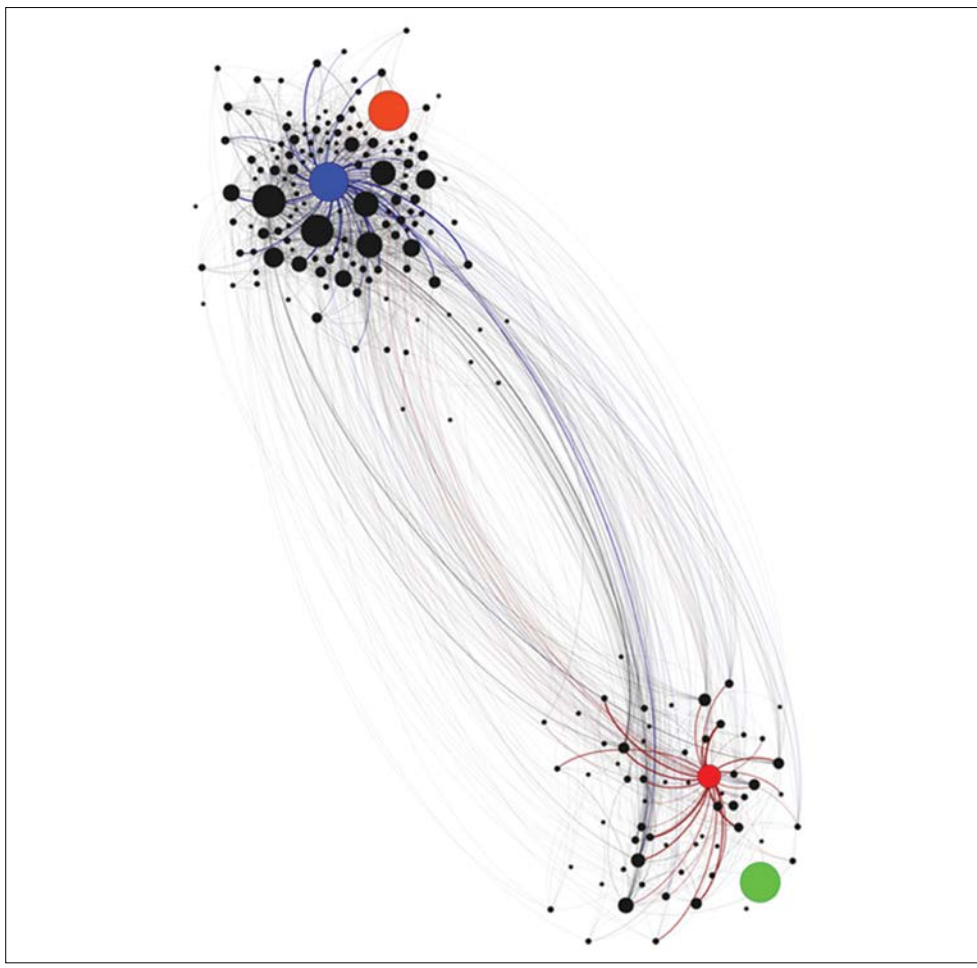
Hu se kraggebaseerde uitlegalgoritme volg in die voetspore van Eades, Kamada en Kawai, en Fruchterman en Reingold, maar stel 'n vinniger berekening voor (wat 'n soortgelyke resultaat oplewer binne 'n baie korter tyd en met inagname van 'n baie groter hoeveelheid entiteite). OpenOrd volg ook in die voetspore van spesifiek Fruchterman en Reingold, maar is ontwikkel om visualiserings van netwerke met meer as 100 000 entiteite (Martin et al. 2011:2) te behartig, en kan soos Hu se uitlegalgoritme ook aangewend word om die sentrale rolspelers in 'n groot netwerk te identifiseer. Figuur 18 demonstreer weer die internasionale wapenhandelnetwerk, waar A) deur middel van die Fruchterman en Reingold uitlegalgoritme voorgestel word, B) deur middel van Hu, en C) deur middel van OpenOrd.



Figuur 18. 'n Vergelyking van uitlegalgoritmes

Hier kan gesien word dat alhoewel die uitlegte beduidend verskil, dieselfde rolspelers altyd onderskeidelik binne die kern of op die periferie van die netwerk geposisioneer word. Vir 'n netwerk van hierdie grootte ondervind Fruchterman en Reingold se uitlegalgoritme nie probleme nie (daar is slegs 201 lande en 1 390 skakels hierby betrokke), maar wanneer groter netwerke ter sprake is, kan die visualisering van die netwerk ure neem.

OpenOrd het die vermoë om só gestel te word dat dit groeperings duideliker kan uitlig. Neem as voorbeeld weer die internasionale wapenhandelnetwerk, wat in Figuur 19 voorgestel word op sodanige wyse dat die groeperings beklemtoon word.



Figuur 19. Groeperings in die internasionale wapenhandelnetwerk

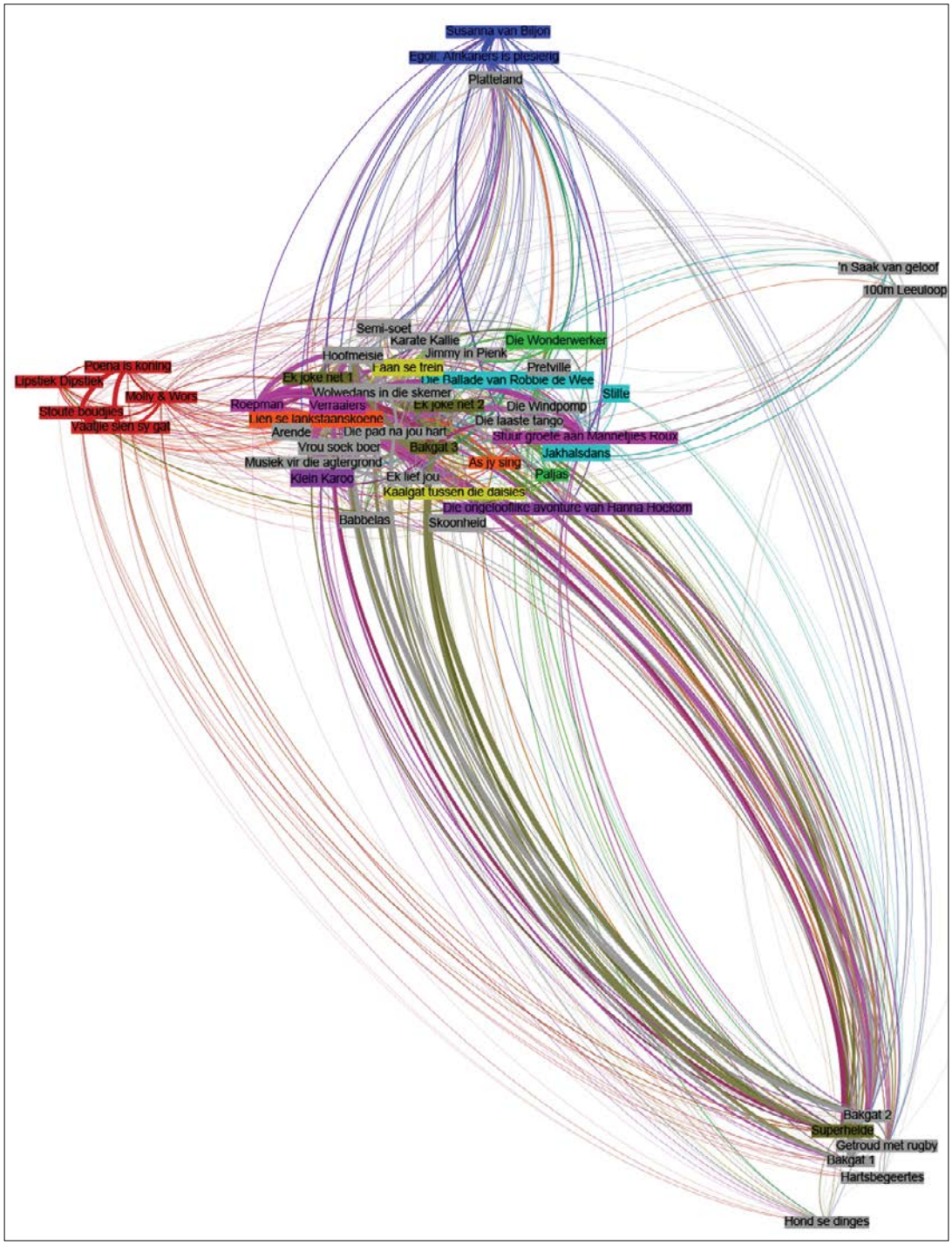
Die oos/wes digotomie is nou nóg duideliker sigbaar, met Suid-Afrika weer eens duidelik in die kamp van die VSA, terwyl die ANC nóg duideliker in die groepering rondom die Sowjetunie geplaas word. Dié funksie van OpenOrd skep verdere ontdekkingsmoontlikhede as wat met Fruchterman en Reingold of Hu die geval is.

Neem byvoorbeeld die voorstelling van die Afrikaanse filmindustrie in Figuur 20: André Odendaal se films is in oranje, Bromley Cawood in blou, Darrell Roodt in seegroen, Katinka Heyns in groen, Koos Roets in geel, Paul Eilers in pienk, Regardt van der Bergh in pers, Stefan Niewoudt in kakie, en Willie Esterhuizen in rooi. Wanneer 'n OpenOrd uitlegalgoritme gebruik word, word films saam gegroepeer waaraan baie van dieselfde mense gewerk het.

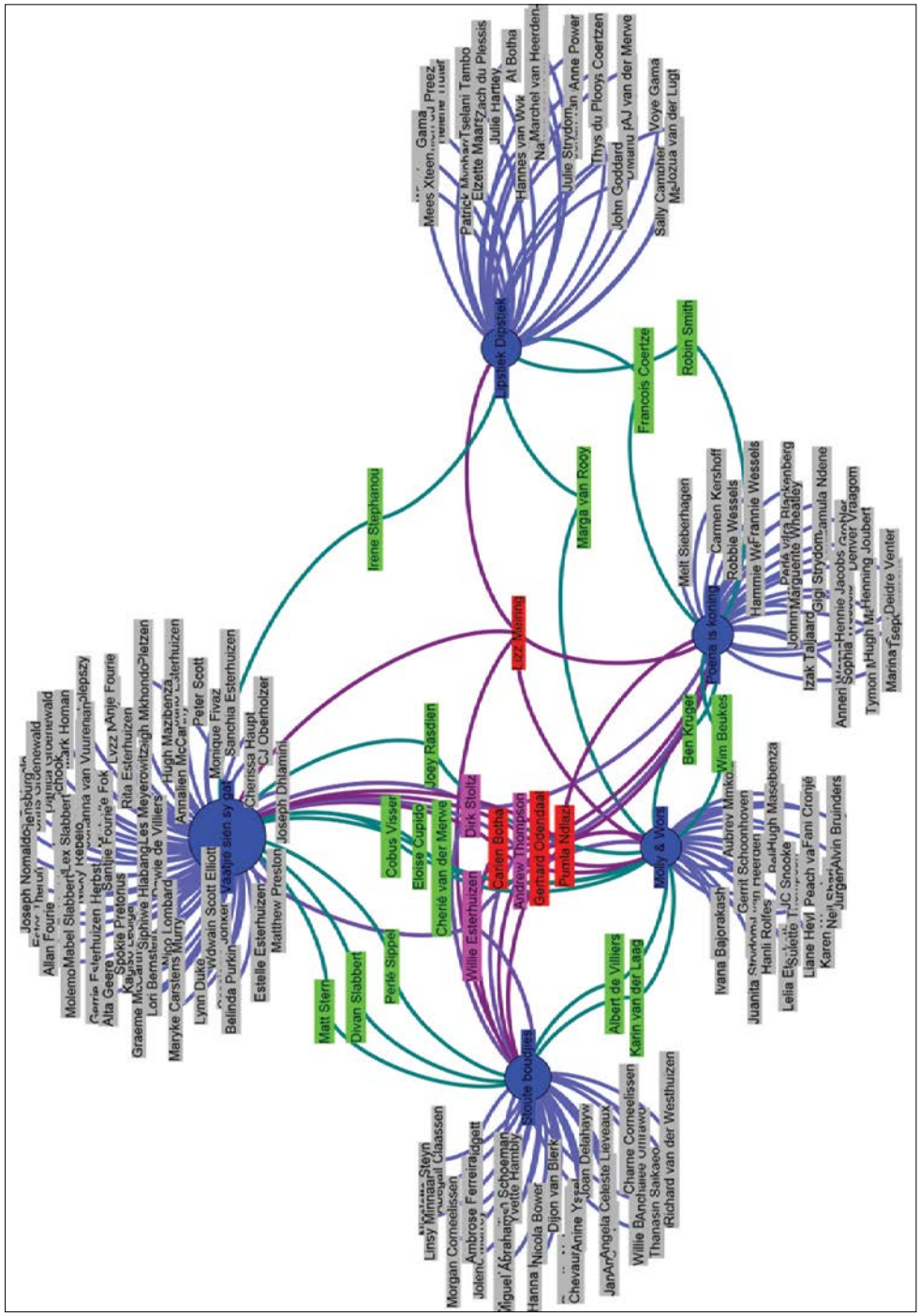
Die groeperings wys dat dieselfde mense by hierdie films betrokke was, en dit is opvallend dat daar 'n groepering ontstaan tussen die Willie Esterhuizen-films wat hier in rooi aangedui is (*Vaatjie sien sy gat, Lipstiek Dipstiek, Stoute Boudjies, Molly en Wors,* en *Poena is Koning*), asook tussen die films wat deur Bromley Cawood geregisseer is en hier in blou aangedui is (*Susanna van Biljon* en *Egoli*). Hierdie groeperings dui daarop dat sommige rolprentmakers gereeld van dieselfde akteurs gebruikmaak, veral in die geval van Willie Esterhuizen en Bromley Cawood. So 'n visuele voorstelling kan die navorser lei om dieper na sy onderwerp te kyk: Wie figureer in die verskillende films van 'n filmmaker? Die grafiek in Figuur 21 stel byvoorbeeld die filmakteurnetwerk in Willie Esterhuizen se films voor, met akteurs wat in vier van sy films gespeel het in rooi, dié wat in drie gespeel het in pienk, en dié wat in twee gespeel het in groen.

Lizz Meiring, Carien Botha, Gerhard Odendaal en Pumla Ndlazi is hiervolgens die akteurs wat tot die grootste mate daarvoor verantwoordelik is dat Willie Esterhuizen se films 'n duidelik identifiseerbare groepering in die Afrikaanse filmnetwerk vorm.

OpenOrd is ontwikkel om parallel oor verskillende verwerkers gebruik te word, 'n beginsel wat onderliggend is aan Hadoop en MapReduce en een van die oplossings vir die grootdataprobleem verteenwoordig (sien Hoofstuk 6). Sodoende kan groter verwerkingskrag aangewend word as wat 'n enkele verwerker kan lewer, wat beteken dat nóg groter netwerke visueel voorgestel kan word, en boonop binne 'n beter tydsraamwerk.

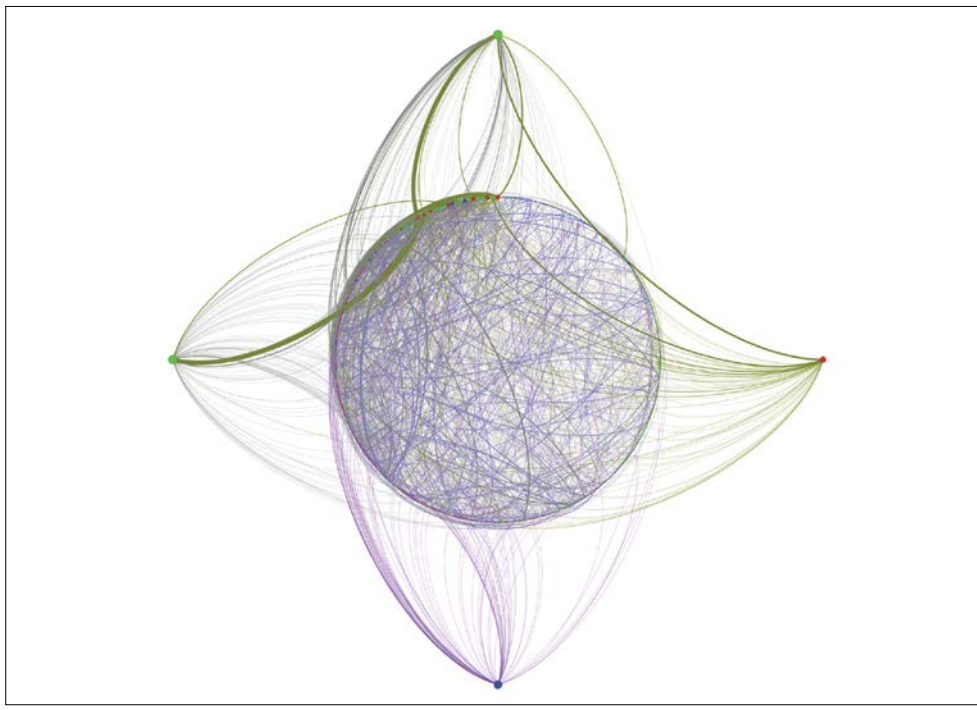


Figuur 20. Die hedendaagse Afrikaanse filmindustrie



Figuur 21. Die filmakteurnetwerk van Willie Esterhuizen se films

Kraggebaseerde uitlegte is natuurlik nie die enigste manier om 'n netwerk voor te stel nie, en ander uitlegte kan ook aangewend word om 'n netwerk voor te stel wat ander inligting oordra. In Senekal (2014:103) word gebruikgemaak van 'n sirkeluitleg om die skakels tussen persone wat betrokke was by die verregse Vaaldam-komplot uit te lig, juis omdat 'n kraggebaseerde uitleg nie geskik was om die nodige inligting oor te dra nie. So kan die navorser byvoorbeeld ook 'n dubbelsirkel-uitleg aanwend om die belangrikste rolspelers in 'n groot netwerk uit te lig. Die grafiek in Figuur 22 dui die hedendaagse Afrikaanse poësiesisteem, soos bespreek in Senekal (2013; 2014) aan. Die aktiefste rolspelers (Joan Hambidge, Bernard Odendaal, Protea Boekhuis en www.litnet.co.za) word aan die buitekant aangedui.



Figuur 22. Die hedendaagse Afrikaanse poësiesisteem in 'n dubbelsirkel-uitleg

Let op die verskillende kleure: Groen dui in hierdie geval op mense (digters sowel as kritici en letterkundiges), blou op uitgewerye, en rooi op publikasieplatforms. Deur verskillende kleure toe te wys aan verskillende soorte entiteite kan uitlegte ook verhelder word om makliker te sien hoe 'n netwerk geskakel is. In hierdie geval is die

netwerk te groot om die titels van entiteite op die bladsy voor te stel, maar vir die navorser wat besig is om met sy datastel om te gaan, is dit natuurlik nie 'n probleem nie.

5.7 Navorsing oor netwerke binne die geesteswetenskappe

Veral binne die geesteswetenskappe beskik die netwerkteorie oor die potensiaal om brûe tussen dissiplines te bou. Die teorie het immers gedeeltelik ontwikkel as gevolg van die insigte van antropoloë soos Kurt Lewin (1951) en Alex Bavelas (1948), en sosioloë soos Jacob Moreno (1934), Stanley Milgram (1967) en Mark Granovetter (1973). Selfs die fisikus Duncan Watts beskou homself tans as 'n sosioloog.

Veral binne die sosiologie en antropologie is die aantal studies wat reeds met behulp van SNA onderneem is te veel om te lys. Een interessante studie wat wel uitgesonder kan word, is die genoemde studie van Christakis en Fowler (2007) wat ondersoek ingestel het na die verspreiding van vetsug oor sosiale netwerke. Dié is 'n oorlangse studie wat 32 jaar se data ondersoek het, en bevind het dat vetsug as 't ware 'aansteeklik' is in die sin dat mense gewigsprobleme ontwikkel wanneer ander mense in hul sosiale netwerke gewigsprobleme ontwikkel. In 'n latere boek (2010) vat die outeurs hul verskeie navorsingsprojekte saam en dui daarop dat ook geluk, depressie, selfmoord en om op te hou rook oor sosiale netwerke versprei.[42]

Rakende taal self is daar reeds 'n verskeidenheid studies gepubliseer oor die struktuur van Engels as komplekse netwerk[43] (Beckner et al. 2009; Ferrer i Cancho & Solé 2001; Dorogovtsev & Mendes 2001; Masucci & Rodgers 2006; Smith, Brighton & Kirby 2003; Motter et al. 2002; Solé, Corominas-Murtra, Valverde & Steels 2010). Hierdie studies ondersoek die verbande tussen woorde in 'n taal, hetsy semanties of sintakties, soos hierbo met betrekking tot "Die stem" geïllustreer is. 'n Soortgelyke studie is nog nie in Afrikaans onderneem nie.

42 Veral die vetsug-studie het berug geword, en William Shatner se karakter Denny Crane op die televisieprogram *Boston Legal* het gedreig om sy oorgewig assistent af te dank as gevolg van die gesondheidsrisiko wat sy vir hom ingehou het, terwyl Jay Leno die studie as 'n grap gebruik het op *The Tonight Show* (Barabási 2011:232). Naas Milgram se klein-wêreld-studie is dié dus ook 'n wetenskaplike studie wat die wêreld van populêre diskoers betree het.

43 Hierdie studies is wel binne fisika onderneem, maar die onderwerp val duidelik binne die geesteswetenskappe. Soos voorheen genoem is dié tendens juis 'n probleem, aangesien die geesteswetenskappe opsy geskuif word deur die fisika in die bestudering van die menslike beweegruimte.

In die letterkunde is SNA toegepas binne die veld- en sisteemteorie (De Nooy 1991, 1993, 2003; Senekal 2013, 2014), en Amancio, Oliveira en Costa (2012) het ook netwerkberekenings gebruik om literêre bewegings tussen 1590 en 1922 te identifiseer. Jockers (2013) gebruik Gephi om ondersoek in te stel na hoe literêre invloed versprei. 'n Ander noemenswaardige studie is dié van Park, Kim, Hwang en Cho (2013), wat 'n statistiese ontleding onderneem het om die hoofkarakters in 'n aantal tekste te identifiseer. Ook is daar 'n verskeidenheid studies oor die sosiale netwerke van karakters in literêre werke, onder andere dié van Shakespeare, gedoen (Stiller, Nettle & Dunbar 2003; Stiller & Hudson 2005). In Afrikaans is die familiebande van karakters in Etienne van Heerden se *Toorberg* al só ondersoek (Senekal 2013).

In die geskiedenis is Padgett en Ansell (1993) se studie van sosiale netwerke en die Medici familie in die 16de eeu veral van belang. 'n Mens sou ook die wye verskeidenheid studies van terroristenetwerke onder geskiedenis of politieke wetenskap kon tel, byvoorbeeld Krebs (2002), Rodriguez (2005), Koschade (2007), Henke (2009), Aghakhani, Dawoud, Alhajj en Rokne (2011), en Wiil, Gniadek, en Memon (2011). In 'n Suid-Afrikaanse opset ondersoek Senekal (2014c) die verregse Vaaldam-komplot, en stel voor dat 'n mens die Boeremag, sowel as Islamitiese terroriste, se netwerke só sou kon ondersoek (laasgenoemde is ook belangrik in 'n Suid-Afrikaanse konteks).

Ook in antieke kultuurstudies is die netwerkteorie reeds toegepas, byvoorbeeld deur Alexander en Danowski (1990), Malkin (2011), Malkin, Constantakopoulou en Panagopoulou (2011), Cline (2012) en Broekaert (2013). Cline is veral 'n interessante figuur omdat sy vir 'n lang tyd ná 11 September 2001 in die militêre intelligensiegemeenskap gewerk het voor sy teruggekeer het na die akademie, en toe die netwerkteorie binne antieke kultuurstudies begin toepas het. Weer eens dui haar loopbaan op hoe vervleg die netwerkteorie met militêre intelligensie is.

Filmakteurnetwerke is ook al deur middel van die netwerkteorie ondersoek, en is sedert Watts en Strogatz (1998) byna 'n klassieke toepassing daarvan. In die meeste van hierdie studies word die netwerke van akteurs ondersoek, en data kan enorme proporsies aanneem – in Guillaume en Latapy (2006) word 'n internasionale filmakteurnetwerk van 392 340 akteurs en hul 15 038 083 onderlinge verbintenisse ondersoek. In Afrikaans stel Senekal en J.-A. Stemmet (2014) ondersoek in na die posisie van Jamie Uys in die Afrikaanse filmindustrie, en Senekal (2014a) ondersoek Pierre de Wet se medewerkers en posisie in die Afrikaanse rolprentbedryf.

Ook in die opvoedkunde het netwerkontledings neerslag gevind; Badge, Saunders, en Cann (2012) gebruik byvoorbeeld Gephi om studente-interaksies te bestudeer. SNA vorm ook deel van RGK-studies.

5.8 Gevolgtrekking

'n Groot verskeidenheid netwerke kan met behulp van netwerkontleding voorgestel word, maar grafiese voorstellings is slegs 'n hulpmiddel: die netwerkteorie is 'n ontledingsinstrument én 'n teoretiese raamwerk. 'n Groot hoeveelheid navorsing het reeds aangetoon dat al vier soorte netwerke wat hierbo bespreek is strukturele kenmerke deel, wat beteken dat ontdekkings wat op een terrein gemaak word ook op ander dissiplines van toepassing is: aldus Strogatz (2004:256) het netwerke dieselfde skelet wanneer die vlees verwyder word.

Die netwerkteorie is in 'n sekere opsig 'n bloedjong benadering, en daarom is daar heelwat verdere studie wat hiermee onderneem kan word. Dié benaderingswyse is by uitstek 'n tegnologiese een wat sterk op visualiserings steun, is deeglik wetenskaplik bewys én het toepassingsmoontlikhede in die nie-akademiese leefwêreld (veral in militêre intelligensie) gevind. As benaderingswyse binne die teorie van komplekse sisteme en grootdata is dit ook 'n teoretiese raamwerk wat oor die afgelope dekade en 'n half aan die voorpunt van die wetenskap gestaan het. Een nadeel van 'n netwerkbenadering is dat die meeste programmatuur slegs met gestruktureerde data kan omgaan, wat beteken dat die verwerkingsfase van 'n navorsingsprojek aansienlik vergroot word wanneer ongestruktureerde data in 'n gestruktureerde formaat omgeskakel moet word.

Hoofstuk 6

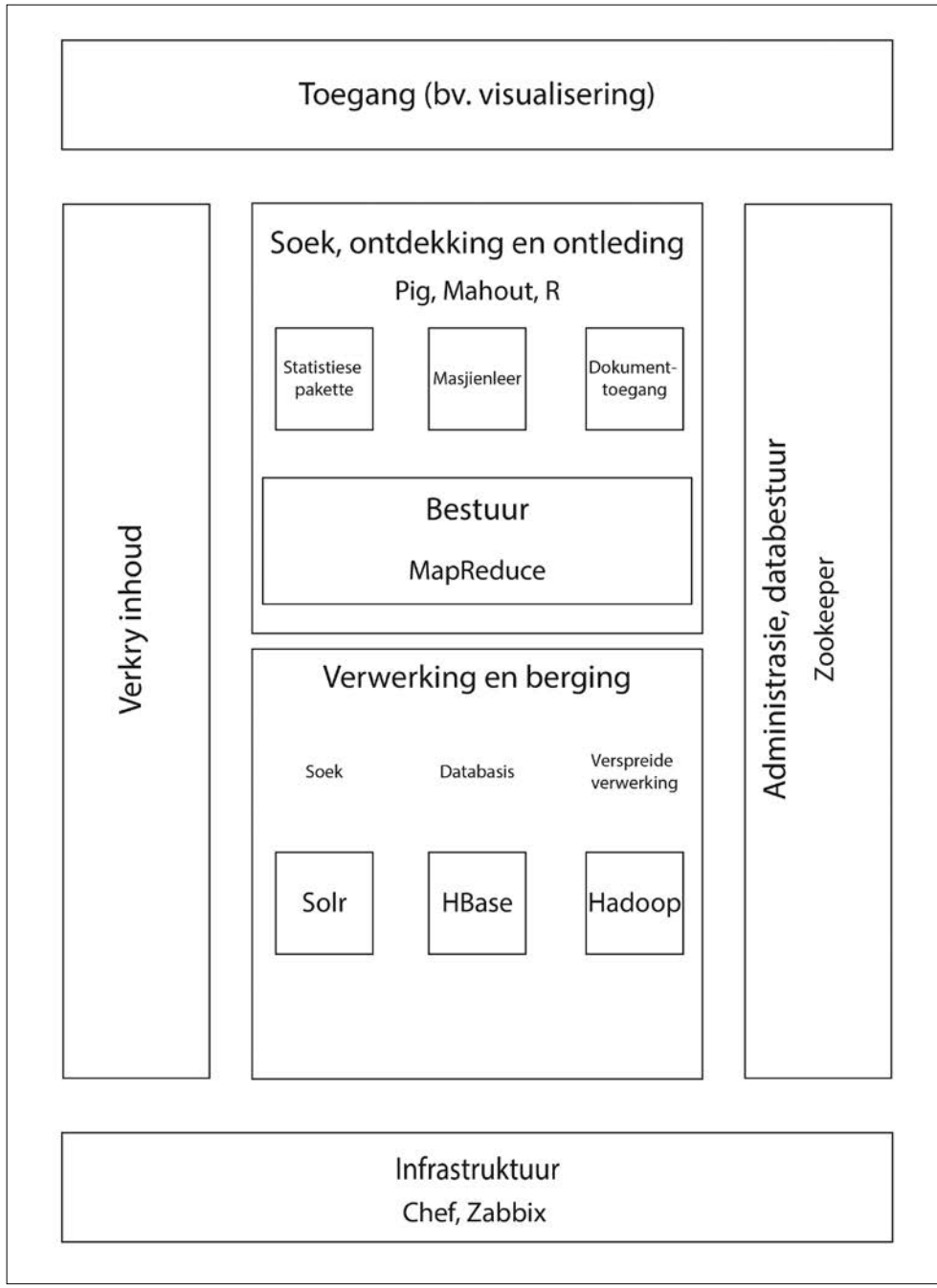
Grootdata versameling, verwerking en ontleding

6.1 Inleiding

Die voorafgaande hoofstukke het ’n middeweg gesoek vir die hantering van grootdata, onder andere deur die versamelingsproses binne die Inligtingsera te bespreek, die ontledingsproses deur groter ongestruktureerde datastelle te ontleed met behulp van kwalitatiewe rekenaarprogrammatuur, en deur met gestruktureerde datastelle binne die netwerkteorie om te gaan. Die huidige hoofstuk bespreek egter ’n meer radikale grootdatabenadering, aangesien ’n mens nie ’n oorsig kan bied oor grootdata sonder om dié veld aan te raak nie. Deurgaans is ’n poging aangewend om aan die een kant agtergrond te verskaf, en aan die ander kant haalbare inisiatiewe te bespreek: die navorser binne die geesteswetenskappe sal waarskynlik nie sommer nodig hê om petagreepdata deur parallelle verwerkers in reële tyd te verwerk nie, maar moet tog kennis neem van wat binne die veld van grootdata gebeur.

Grootdataprogrammatuur wissel van peperduur na gratis. Die Apache Software Foundation verskaf oopbronprogrammatuur, hoofsaaklik by Yahoo!, Google, LinkedIn en Facebook ontwikkel, wat ’n integrale deel van grootdatabestuur geword het. Dié programmatuur verg egter kundigheid, en in ’n poging om programmatuur meer gebruikersvriendelik te maak het ’n aantal private ondernemings oor die afgelope paar jaar hul eie soortgelyke produkte ontwikkel, en dié is gewoonlik baie duur. Die huidige hoofstuk verskaf slegs ’n oorsig oor wat beskikbaar is en wat daarmee vermag kan word, en ons volg outeurs soos Chen, Mao, Zhang en Leung (2014), Kambatla et al. (2014) en Krishnan (2013) deur te fokus op oopbronprogrammatuur.

Die komende bespreking kan verwarrend wees as gevolg van die kompleksiteit van grootdataprogrammatuur, en daarom is dit sinvol om eers na ’n diagrammatiese voorstelling (Figuur 23) van tipiese grootdata-infrastruktuur, soos aangepas uit Ingersoll (2012:6), te kyk.



Figuur 23. Grootdata infrastruktuur

Dié kategorieë is egter nie onbeweeglik nie, en Solr kan byvoorbeeld beide databerging as soekfunksies vermag, terwyl R beide 'n statistiese ontleding as 'n visualisering kan doen. In die komende bespreking sal meer besonderhede oor sommige van hierdie programme verskaf word.

6.2 Versameling

Daar is altyd plek in die navorsingproses vir bogenoemde aktiewe en passiewe soekstrategieë, maar om groot datastelle van miljoene dokumente te versamel word 'n geoutomatiseerde proses benodig. Hiervoor bestaan daar produkte soos Apache Nutch, wat sedert 2005 deur Doug Cutting en ander ontwikkel is. Nutch is 'n sogenaamde webkruiper wat die internet vir data deursoek, ontwerp is om op die Apache Hadoop platform te werk, en biljoene webwerwe kan indekseer (Mattmann & Zitting 2011:162-163). Soos Hadoop is dit oopbronprogrammatuur, en kan dit inkoppel met Apache Tika en Solr (sien hieronder) om 'n data-ontledingsstroom te vorm, maar Nutch kan ook self natuurlike taalverwerking (Natural Language Processing of NLP)[44] en data-ontginning behartig. Nutch kan dus aangewend word om astronomiese datastelle met behulp van die web te versamel (die tegniese aspekte van Nutch is onder andere in 2013 deur Nioche bespreek).

Wanneer groot datastelle versamel is, moet dit natuurlik geberg word met behulp van 'n rekenaarprogram wat groot datastelle kan hanteer, aangesien programmatuur soos Microsoft Access en Excel tekortskiet. NoSQL (Not Only Standard Query Language) (die term is geskep deur Eric Evans van die Apache Foundation) is ontwikkel as oopbronprogrammatuur om verskeie datatipes te hanteer (Krishnan 2013:86-87), groter hoeveelhede data te stoor as wat met SQL en soortgelyke tegnologie moontlik is, en verteenwoordig tans die 'kern' van grootdataberging (Chen, Mao & Liu 2014:186). Cassandra is ook vir dié doel by Facebook ontwikkel, word gesien as 'n vorm van NoSQL, en is deels gebaseer op Google se BigTable, wat 'n soortgelyke databasis is (Krishnan 2013:88-96; Chen et al. 2014:41). Facebook stoor gebruikersinligting soos foto's en boodskappe in Cassandra (Kambatla et al. 2014:2566), en dié program is ook in gebruik by Twitter. Apache stel sedert 2008 'n oopbron-weergawe van dié program beskikbaar.

Apache HBase (Being Available and Same Everywhere) is 'n skaalbare, verspreide databasis wat gestruktureerde data in groot tabelle kan berg – selfs tabelle met biljoene

44 Natuurlike taalverwerking is 'n veld binne rekenaarwetenskap wat ondersoek instel na hoe rekenaars met menslike taal kan omgaan, beide in die verstaan van taal en in die generering daarvan, byvoorbeeld sentiment-ontleding van boodskappe op Twitter.

reëls en miljoene kolomme (Loukides 2010:5). Dit word geklassifiseer as 'n NoSQL databasis en is ook gebaseer op Google se BigTable (Chen et al. 2014:16; Krishnan 2013:74).

HDFS (Hadoop Distributed File System) is 'n verspreide lêerstelsel wat dokumente van tot petagreepgroottes kan verwerk, en is ook een van die kerns van databerging in Hadoop (Krishnan 2013:54-60, 75; Chen et al. 2014:16). Al hierdie bergingsmetodes word gewoonlik saam aangewend om elkeen se swakpunte uit te skakel.

'n Ander interessante manier om grootdata te stoor is die grafiekdatabasis. Dit is veral die gevolg van sosiale media en die hedendaagse wêreld se besef dat niks geïsoleer is nie, en stoor data in 'n vorm soortgelyk aan 'n netwerk, met elke objek as 'n nodus en 'n skakel tussen verskillende verwante nodusse. Grafiekdatabasisprogrammatuur sluit in Neo4J, infiniteGraph, GraphDB, en AllegroGraph (Krishnan 2013:97), en kan gewoonlik ook datastelle van petagreepgroottes stoor en inligting blitsvinnig daaruit herwin. Neo4j is oopbronprogrammatuur en werk saam met die programmeringstaal Cypher.

6.3 Verwerking

Die verwerking van grootdatastelle sal geheueprobleme (en gevolglike lae spoed) vir 'n rekenaar veroorsaak, en om hierdie probleem te oorkom is die Hadoop platform deur Doug Cutting en Mike Cafarella in samewerking met Yahoo! geskep (Chen et al. 2014:16; Krishnan 2013:53; Mattmann & Zitting 2011:16). Sedert dit in 2006 vrygestel is maak dit deel uit van die gratis produkte wat deur die Apache Software Foundation gebied word, en word onder andere gebruik deur Yahoo! en Facebook (laasgenoemde beweer hul implementering van Hadoop kan 100 petagrepe se data verwerk (Chen et al. 2014:17; Chen, Mao & Liu 2014:178)). Hadoop versprei data oor verskeie verwerkers en maak dit moontlik om 'n groot hoeveelheid berekeninge sodoende te voltrek, en die verspreiding van data oor verskillende bedieners het die voordeel dat oorbodige duplisering data teen verlies beskerm.[45] Yahoo! gebruik tans 42 000 bedieners vir hierdie doel (Chen et al. 2014:17). Hadoop het reeds só gewild geword dat daar verskeie alternatiewe Hadoops bestaan, soos Cloudera Hadoop. Kambatla et al. (2014:2567) skryf dat Hadoop binnekort betrokke sal wees by die helfte van die wêreld se data.

Die Hadoop-platform word gebruik om verdere infrastruktuur op te bou, soos deur MapReduce en HDFS. Hadoop MapReduce is 'n sagteware-raamwerk vir die verspreide verwerking van groot datastelle. Dit is aanvanklik geskep deur Google,

45 Die details van hoe die Hadoop platform werk, word bespreek in Krishnan (2013:54 e.v.) en Chen et al. (2014:16).

en bestaan uit twee komponente, naamlik kartering en skeiding (laasgenoemde verdeel data in kleiner pakkies sodat dit oor verskeie verwerkers verwerk kan word). Sedert 2012 is daar ook 'n nuwe weergawe van MapReduce genaamd YARN (Krishnan 2013:60-69; Chen et al. 2014:16). Apache ZooKeeper is 'n hoëspoed produk vir verspreide verwerkings wat onder andere koördinering vir die verskeie onderlinge produkte verskaf (Krishnan 2013:69-72), byvoorbeeld deur Hadoop met MapReduce, HDFS en HBase te laat skakel. Die programmeringstale Hive, Pig en Python, wat almal ook sterk met grootdatametodes geassosieer word, werk ook op Hadoop. Apache Hive is 'n datapakhuisinfrastruktuur wat data-opsomming bied en ad hoc soektogte fasiliteer (dit is by Facebook ontwikkel) (Krishnan 2013:78-82), en Apache Pig (ontwikkel by Yahoo!) is 'n hoëvlak datavloeitaal en uitvoeringsraamwerk vir parallelle verwerking (Krishnan 2013:72-74). Saam met Google se Sawzall en Microsoft se Scope verskaf dit ook 'n meer gebruikersvriendelike koppelvlak met die res van die Hadoop-sisteem (Chen et al. 2014:46). Pig, Hive en Python is die bekendste programmeringstale in grootdata (Davenport (2014:132).

Tot dusver is programmatuur genoem wat grootdata verkry en bestuur. Daar bestaan ook 'n groot hoeveelheid programmatuur wat data-ontginning en teksontleding kan behartig, byvoorbeeld Apache Mahout, 'n gradeerbare masjienleer- en data-ontginningsplatform (Krishnan 2013:54). Mahout kan trosontledings behartig waar soortgelyke dokumente op grond van hul inhoud saam gegroepeer word (Owen et al. 2012:145 e.v.). Krishnan (2013:240) skryf egter dat dit veral in hierdie fase is dat opgeleide datawetenskaplikes benodig word omdat hierdie 'n uiters komplekse proses is. Mahout word in detail bespreek in Owen et al. (2012).

Apache Solr kan gebruik word vir teksontledings en om spesifieke soektogte na inligting te doen, maar soos Ingersoll (2012:7) aandui kan dit ook tot 'n beperkte mate vir databerging aangewend word, en bemiddel dit ook trosontledings (Grainger & Potter 2014:22). Solr word in detail bespreek in Grainger en Potter (2014).

Apache Tika is in 2006 deur Jerome Charron en Chris Mattmann begin as deel van die uitgebreide Apache Nutch-projek (Mattmann & Zitting 2011:16). Tika kan teks onttrek vanuit gestruktureerde en ongestruktureerde dokumente, en kan dan dié inligting in 'n gestruktureerde formaat berg. Mattmann en Zitting (2011) bespreek die gebruik van Tika vir teksontleding en data-ontginning in detail.

R is 'n ander welbekende oopbrondata-ontginning- en ontledingsprogram wat statistiese ontledings en visualisering kan behartig. Dié program het reeds bykans die standaard in statistiese ontledings geword en is ontwikkel deur Robert Gentleman en Ross Ihaka, wat die program op John Chambers se S gebaseer het. Sedertdien het

'n groot gemeenskap oopbronprogrammeerders bygedra tot die ontwikkeling van R, en daar bestaan tans meer as 5 000 toevoegings tot dié program, sowel as weergawes wat aangekoop kan word (en gevolglik meer gebruikersvriendelik is). R koppel met die hele Apache-infrastruktuur, maar is ook selfstandig, en kan ook skakel met ontledingsprogrammatuur soos Tableau. R word in detail bespreek in Zumel en Mount (2014), en is uniek in die opsig dat daar reeds doelgerigte handleidings geskryf is oor hoe dié program vir navorsingsdoeleindes binne die geesteswetenskappe aangewend kan word. Jockers (2014) verskaf 'n praktiese gids tot die gebruik van hierdie program vir die literatuurstudie, en in die linguistiek fokus Baayen (2008) en Gries (2009) ook spesifiek op dié program. R vereis egter programmeringsvaardighede, maar Jockers, Baayen en Gries verskaf praktiese riglyne wat die beginner in staat stel om dié sleutelprogram te bemeester en die brug tussen letter- en/of taalkunde en datawetenskap oor te steek.

6.4 Ontleding

Ten einde groot hoeveelhede data te ontleed, moet rekenaarprogrammatuur uiteraard ingespan word, en die visualisering van data is 'n belangrike komponent van grootdata-ontledings (Park & Leydesdorff 2013:756; Schöf 2013:9; Loukides 2010:7; Keim, Qu & Ma 2013). Die visualisering van data is absoluut noodsaaklik (Fox & Hendler 2011:706) in die ontleding van komplekse datastelle, en dien die tweeledige doel van die fasilitering van ontledings en die visuele voorstelling van bevindings (Keim, Qu & Ma 2013:50; Agrawal et al. 2011:7).

Visuele ontleding word gedefinieer as die wetenskap van analitiese redenering soos gefasiliteer deur interaktiewe visuele gebruikerskoppelvlakte (National Visualization and Analytics Center 2005:4). Wanneer ons iets verstaan, sê ons in Afrikaans of Engels: 'Ek sien wat jy bedoel' of 'I see what you mean'. Hierdie uitdrukking is 'n manifestasie van die ingebore verband wat die mens lê tussen visie, visualisering, en ons redenasieprosesse (National Visualization and Analytics Center 2005:4; Jessop 2008:281). Programme wat spesialiseer in visualisering poog om bestaande denkprosesse te ondersteun (wat ook die geval is met netwerkontleding). Die hoofvoordeel van visualisering bo teks is dat dit begrip verbeter en daarom as ontdekkingsinstrument aangewend kan word in verskeie velde binne die geesteswetenskappe.

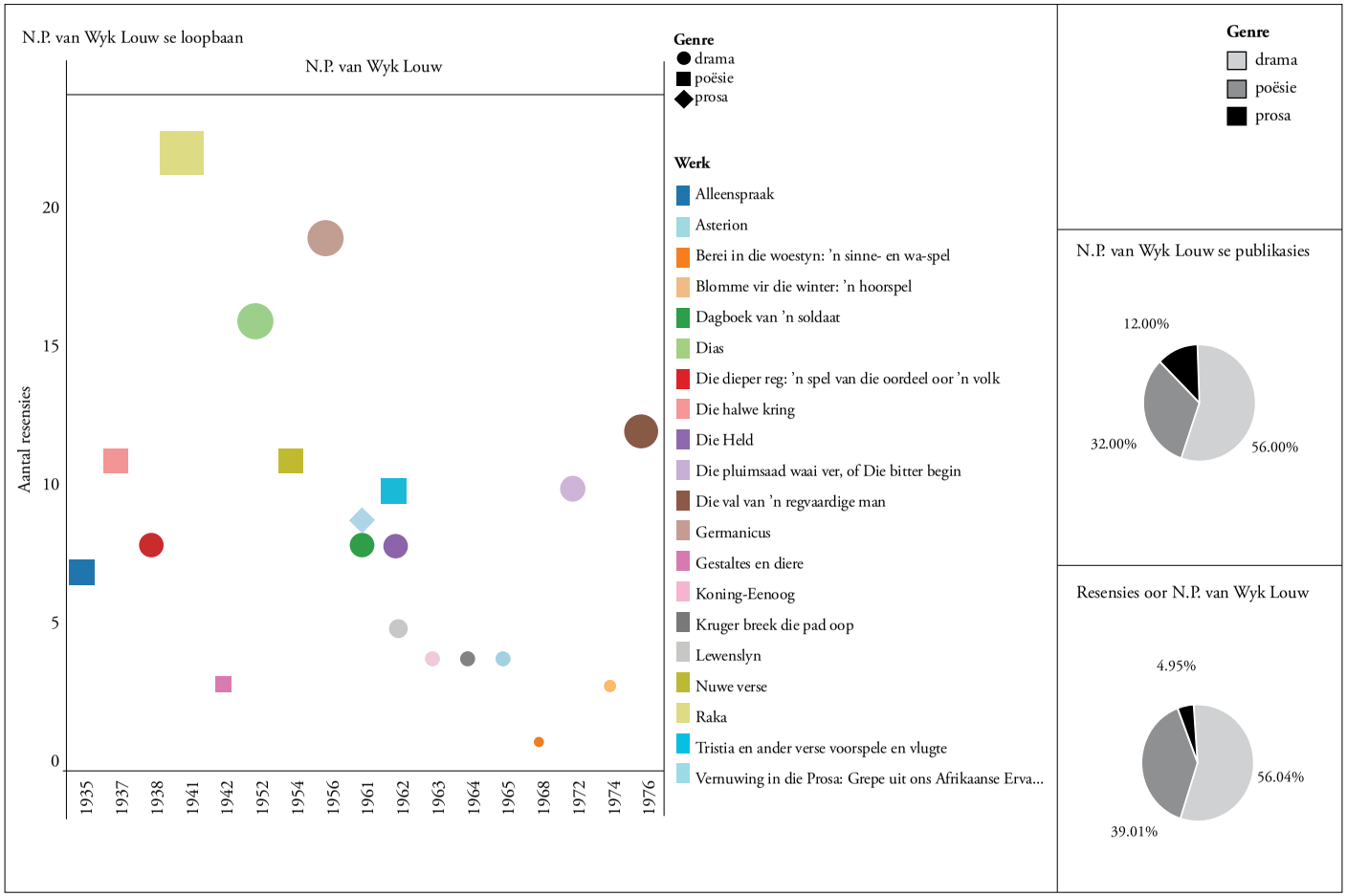
Daar is vele visualiseringsprodukte op die mark. Kirschenbaum (2007:4) noem hoe TIBCO Spotfire, TextArc, en ander al met sukses aangewend is binne die letterkunde, terwyl Athenikos (2009) verduidelik hoe visualisering deur middel van Prefuse kan help met die studie van filosofie. Jockers (2013:15) noem ook TactWeb, TAPoR, MONK, SEASR en DARIAH binne die letterkunde, en het ook bogenoemde gids tot die gebruik

van R geskryf, wat ook tot visualiserings in staat is. Voyant en HyperPro is aanlyn teksontledingsprogramme wat deur Stefan Sinclair en Geoffrey Rockwell ontwikkel is as deel van die Hermeneuti.ca-projek, en voorbeelde van hoe eersgenoemde aangewend kan word in die Afrikaanse letterkunde word aangetoon in Senekal (2012a).

Volgens Lima (2011:12) het die aard van visualiserings verander as gevolg van die huidige wetenskaplike paradigma asook grootdata se klem op omvattendheid. Volgens hom is ouer vorme van visualisering ingebed in die wetenskap se reduksionistiese benadering van vroeër, en visualiserings breek dus ook datastelle af om op hierdie manier sin te maak daarvan. In die huidige paradigma, met die klem op omvattendheid, kompleksiteit en samehang, is visualiserings eerder daarop gemik om verhoudinge, asook die datastel as geheel, te visualiseer om sodoende die huidige wetenskap se klem te versinnebeeld.

Een van die programme wat die voortou neem in die visuele ontleding van grootdatastelle is Tableau. Dié program is ook ontwikkel in samewerking met die VSA se verdedigingsindustrie, maar het 'n verskeidenheid toepassings in die besigheidsektor gevind. eBay gebruik dié program om die toepaslikheid van soekresultate te ontleed (Chen & Zhang 2014:321), en so ook Apple, Google, Microsoft, Walmart, Ferrari, Barclays, Coca-Cola, Toyota, Dell, Vertx, en vele meer. Tableau integreer met Hadoop, Access, Microsoft Excel, Actian Vectorwise, R, FireBird, Claudera Hadoop, Oracle, Splunk en die meeste ander grootdatabergingsinfrastrukture. Neem byvoorbeeld die visualisering van N.P. van Wyk Louw se loopbaan in terme van gepubliseerde werke in Figuur 24, wat met behulp van Tableau gedoen is en gegrond is op die reeds genoemde datastel uit Senekal en Van Aswegen (1980, 1981) en Senekal en Engelbrecht (1984).

Daar is baie inligting in hierdie beeld vervat. Eerstens kan ons in die sektordiagramme onder sien dat die meerderheid van N.P. van Wyk Louw se publikasies in dié datastel dramas was (56%), en dat die meerderheid resensies oor sy werk ook gehandel het oor sy dramas (56,04%), maar dat daar relatief meer oor sy poësie geskryf is as sy prosa: 32% van sy publikasies is poësie, maar 39,01% van resensies oor sy werk handel oor sy poësie. Die groot visualisering dui op wanneer hy werke gepubliseer het, en hier kan gesien word dat sy eerste werk poësie was (*Alleenspraak* 1935), en sy laaste in die datastel 'n drama (*Die val van* 'n *regvaardige man* 1976). Die horisontale as stel datums voor, en die vertikale as en die grootte van simbole die aantal resensies wat oor 'n werk gepubliseer is – hier kan duidelik gesien word dat *Raka* (1941) sy werk is waaroor die meeste kritici geskryf het. Sy werk waaroor die tweede meeste kritici geskryf het, is *Germanicus* (1956), en saam met die feit dat die meeste resensies oor sy dramas gehandel het, moet hierdie as 'n belangrike teks beskou word. In Figuur 25 kan daar gekyk word na die breër geheel van watter Afrikaanse outeurs die meeste resensies oor geskryf is.



Figuur 24. N.P.van Wyk Louw se loopbaan in terme van gepubliseerde werke

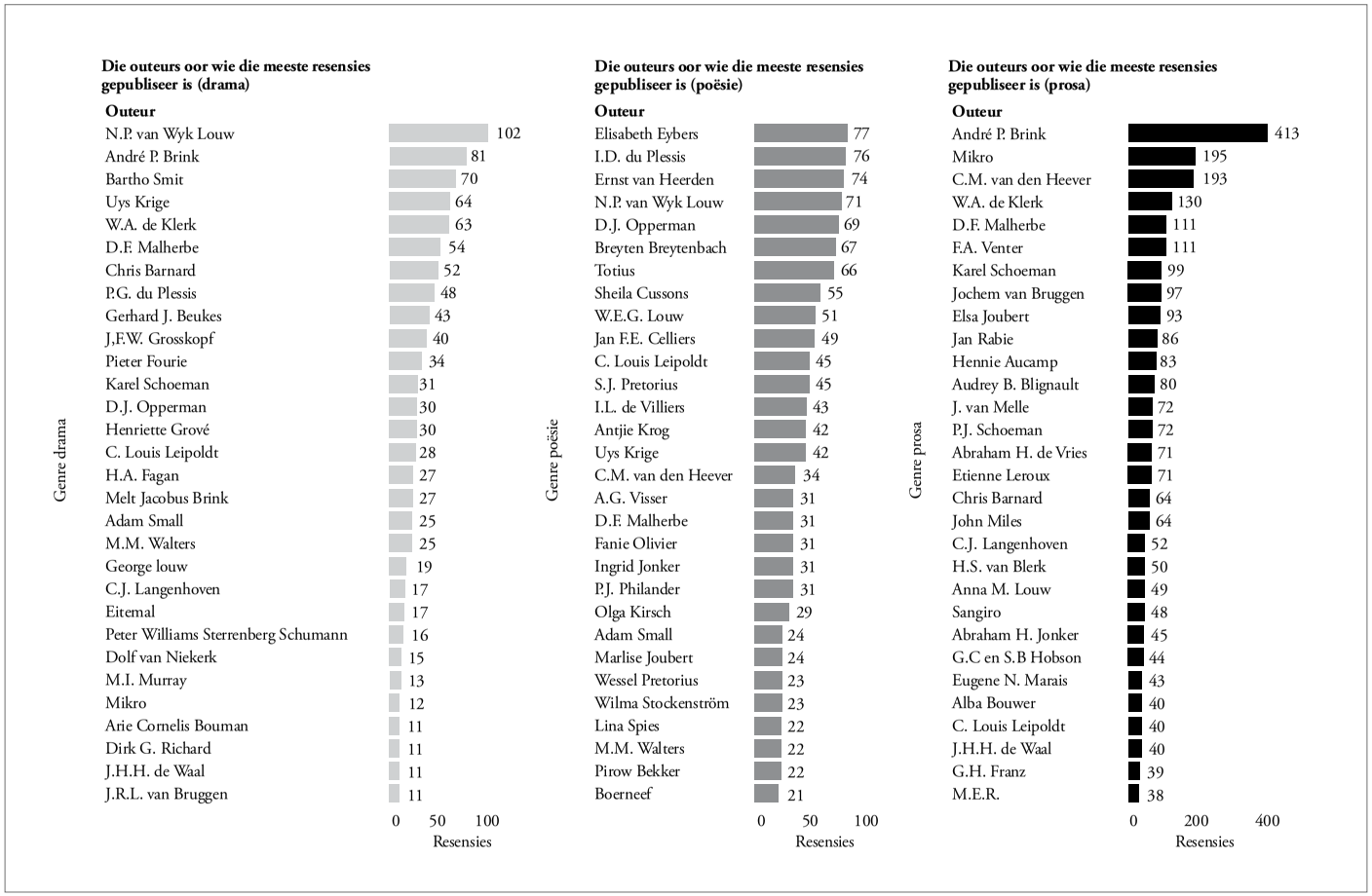
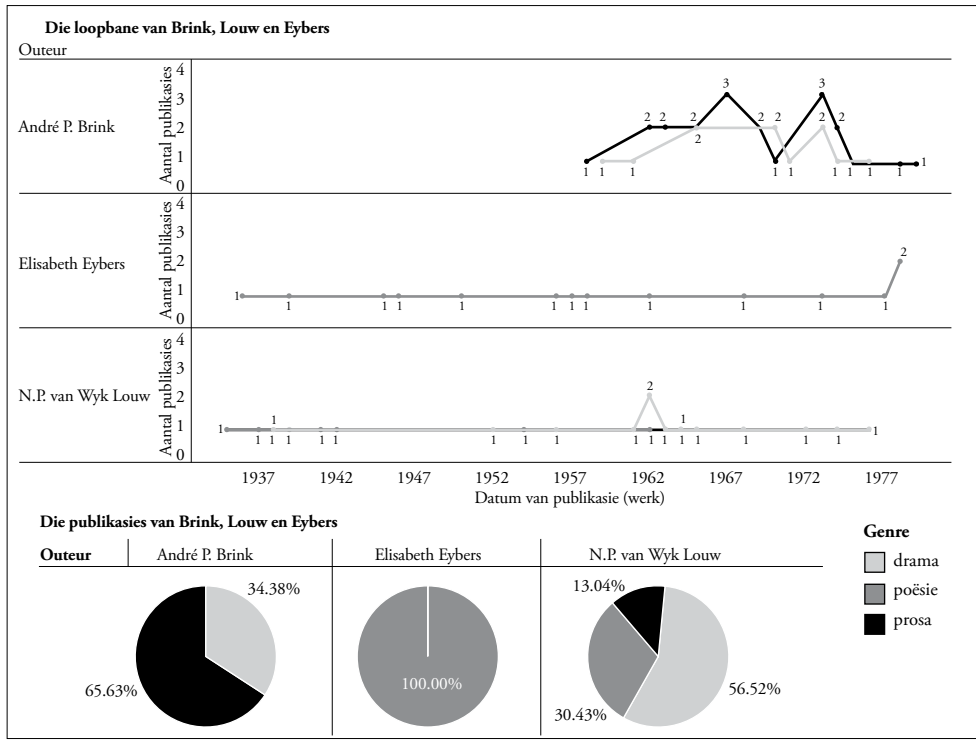


Figure 25. Afrikaanse outeurs oor wie die meeste resensies geskryf is

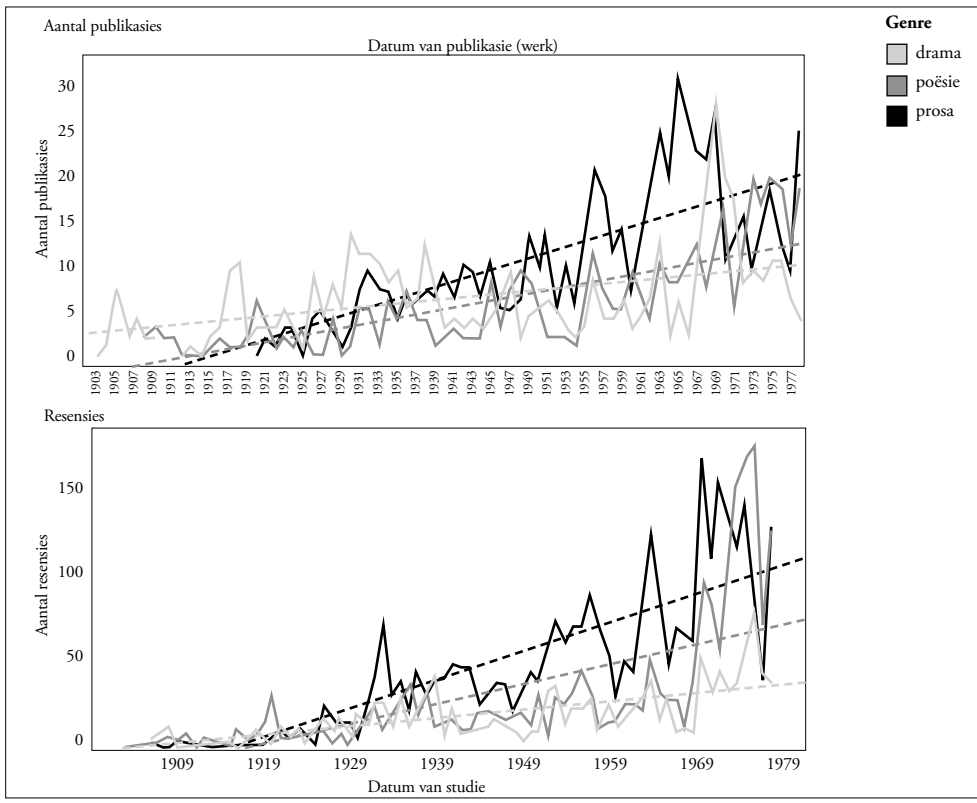
Germanicus het saam met sy ander dramas daartoe bygedra dat N.P. van Wyk Louw in dié datastel die dramaturg is waaroor die grootste aantal resensies handel. Afgesien van netwerkontledings is dít waarna Lima (2011:12) verwys as hy skryf dat visualiserings konteks en verhoudinge beklemtoon. In Figuur 24 is N.P. van Wyk Louw se werke voorgestel in verhouding tot al sy ander werke in die datastel (beide deur die geheelbeeld as deur die twee sektordiagramme), en programmatuur soos Tableau stel 'n mens in staat om die breër geheel te verken soos in Figuur 25. Verdere verkenning word in Figuur 26 gedemonstreer en beantwoord die vraag: Wat is die publikasiepatrone van die outeurs waaroor die meeste resensies geskryf is?



Figuur 26. Brink, Eybers en Louw se publikasiepatrone

Hier kan gesien word dat beide Louw en Eybers lang loopbane gevolg het waarin hulle konstant elke paar jaar 'n boek gepubliseer het, in teenstelling met Brink, wat skielik in die laat vyftigerjare op die literêre toneel verskyn en vinnig 'n groot aantal werke publiseer. Dié skrywers se oeuvres is ook onder op die grafiek opgesom, waar aangetoon word dat Louw die mees veelsydige skrywer van dié drie was, en Eybers die

minste. Hierdie gegewe is natuurlik nie nuus vir ’n kenner nie, maar die punt is dat die visualisering ’n mens in staat stel om ’n maklik verstaanbare opsomming van die data te bied. Ons wonder egter of ’n kenner die opkoms van die prosa as die dominante genre in Afrikaans só duidelik sou kon stel soos in Figuur 27.



Figuur 27. Die opkoms van die prosa

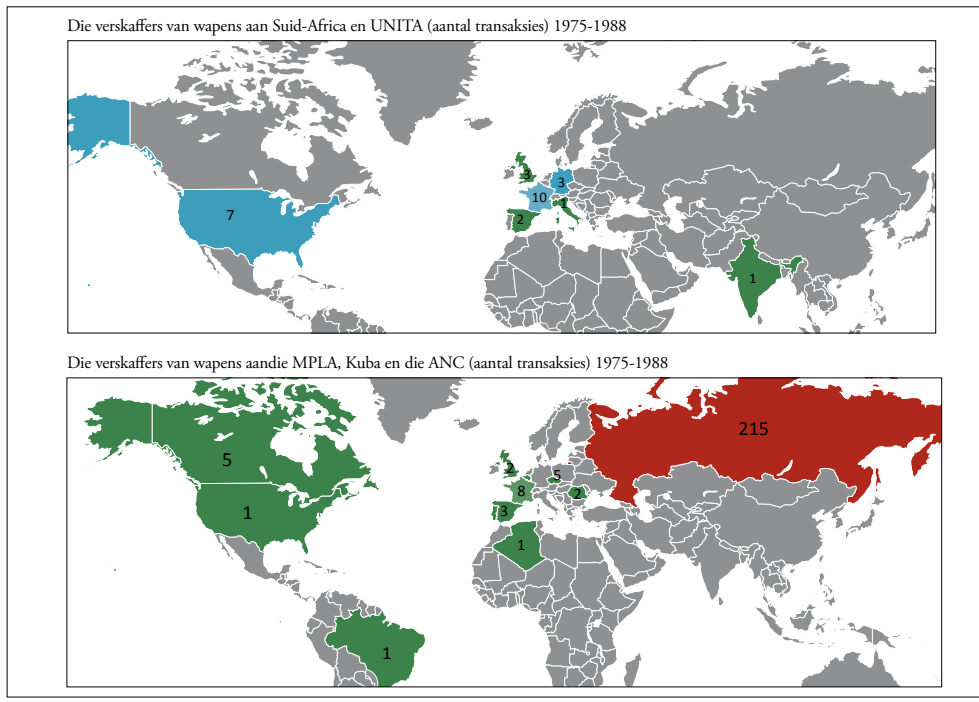
Hier kan gesien word hoe die gemiddelde aantal prosawerke (met stippellyne aangedui) wat gepubliseer is jaarliks sedert 1916 oor die ter sake tydperk (vinniger as die ander genres) toegeneem het, en dat die aandag wat kritici aan werke gegee het nóg vinniger toegeneem het vir die prosa. Deur die bykans 80 jaar hier ondersoek het die prosa sigself duidelik gevestig as die dominante genre in die Afrikaanse literêre diskoers én in die Afrikaanse literêre publikasie-industrie. Hierdie gegewe is iets wat moeilik is om met behulp van ’n ander formaat op ’n enkele bladsy weer te gee.

’n Vorm van visualisering wat onlangs belangriker geword het as gevolg van globalisasie is geografiese visualisering. In die hedendaagse wêreld het nasionale grense vervaag: maatskappye kompeteer internasionaal, inligting versprei oor landsgrense

deur middel van sosiale media en die web, kriminele en terreurnetwerke het ook geglobaliseer, en in die akademie het dit ook al hoe belangriker geword om deel te neem aan die internasionale diskoers binne 'n gegewe veld. Globalisasie raak elke aspek van die wetenskap; reeds in 1987 het Senekal (1987:169) die volgende opgemerk rakende die Afrikaanse literêre sisteem:

> Afrikaanse literêre handelinge bestaan nie in isolasie nie, maar is ten nouste verweef met die internasionale wêreld en sy denke – waarmee dit inderdaad selfs elektronies verbind is. Dit is vandag baie duideliker só as in vorige dekades en toe reeds, van die begin van die Afrikaanse literatuur af, was daar baie sterk import van ander literature na Afrikaans, uit sowel Westerse as uit Afrikatradisies.

Dié stelling dateer uit 'n era kort voor die algemene gebruik van die internet en die web; vandag is die skakeling met die res van die wêreld van nóg groter belang. Om hierdie rede het geografie ook 'n belangrike plek in die visualisering van data, en ook in verskeie dissiplines. Die grafiek in Figuur 28 dui aan waar wapens bekom is wat aangewend is tydens die oorlog in Angola vanaf 1975 tot 1988.[46]



Figuur 28. Wapenverskaffers tydens die oorlog in Angola 1975-1988

46 Data verskry vanaf die Stockholm Institute for Peace Research (SIPRI).

In hierdie visualisering kan die Koueoorlogse digotomie duidelik gesien word: aan die een kant word Suid-Afrika, UNITA (União Nacional para a Independência Total de Angola) en FNLA (Frente de Libertação de Angola) van wapens voorsien deur oorwegend Westerse lande soos die VSA, Frankryk en VK, terwyl die MPLA (Movimento Popular de Libertação de Angola), Kuba en die ANC se grootste aantal wapentransaksies met die Sowjetunie is. Daar is egter ook uitsonderings te bespeur: die VSA, VK, Frankryk en ander Westerse lande het óók wapens aan die Kommunistiese kant voorsien, maar tot 'n mindere mate. Só 'n visualisering is makliker om te verstaan as 'n eenvoudige tabel, en laat nie alleen die navorser toe om sy bevindinge duideliker oor te dra nie, maar ook om sy data beter te verken – 'n mens kan ook afboor na 'n enkele land om in meer detail te sien watter wapensisteme betrokke was by 'n enkele transaksie.

6.5 Gevolgtrekking

Hierdie hoofstuk het 'n kort oorsig gebied van die rekenaarprogrammatuur wat beskikbaar is en aangewend word in die versameling, verwerking, berging en ontleding van grootdata. Die goeie nuus is dat rekenaarprogrammatuur in die toekoms toenemend gebruikersvriendelik sal word; Actian se benadering word verwoord as "big data for the rest of us". Hul oogmerk is om meer gebruikersvriendelike programmatuur te ontwerp om die tekort aan datawetenskaplikes die hoof te bied, die leek toe te laat om om te gaan met sy data, en grootdata-ontledings bekostigbaar beskikbaar te stel. Tableau behoort ook in dieselfde sin genoem te word, aangesien hul program bekostigbaar (selfs gratis vir studente) en uiters gebruikersvriendelik is. Die toekoms van grootdata-ontledings lyk dus heel rooskleurig.

Rekenaarprogrammatuur wat spesifiek vir die ontleding van grootdata geskep is sluit in IBM InfoSphere BigInsights, Kognitio, Ayasdi, SAS Data Integration Studio, Tableau en Actian. Rekenaarprogrammatuur wat binne die veld van militêre intelligensie ontwikkel is en spesifiek toegespits is op grootdata sluit in Starlight VIS en Palantir, maar natuurlik het hierdie programmatuur ook toepassings buite militêre intelligensie.

Slot

Inligtingstegnologie is heelwat dieper ingegrawe binne die navorsingskonteks as wat in hierdie boek bespreek kon word: woordverwerkingsprogramme soos Microsoft Word, sowel as Microsoft Powerpoint as medium vir die verspreiding van bevindinge, is 'n alledaagse realiteit. Selfs hierdie programme word selde optimaal benut: min navorsers gebruik byvoorbeeld Microsoft Word se elektroniese inhoudsopgawe- of bibliografiese verwysingsfunksies, en daar bestaan nog vele ander hulpmiddels wat die navorser se taak kan bespoedig en vergemaklik, soos deur Raubenheimer (2012) uiteengesit. Hierdie boek maak geen bewering dat dit omvattend kan wees nie: dit verskaf bloot 'n oorsig oor wat tans met inligtingstegnologie in die wetenskap gedoen word.

'n Tema wat 'n mens gereeld in die gebruik van data en inligtingstegnologie hoor, is verwysings na netwerke. Neo4j, Gephi, Sentinel Visualizer, Ayasdi, Palantir en NVivo (waar temas as nodusse gekodeer word), benut almal die konsep van 'n netwerk. Dít is nie toevallig nie: die hedendaagse wêreld se interafhanklikheid, en die opkoms van die web, internet en sosiale media, het 'n groter besef van die belangrikheid van konneksies tuisgebring wat ook neerslag vind in die wetenskap. Steven Strogatz (2004:230) skryf dat die wetenskap self die tydsgees reflekteer. Ons leef in 'n wêreld waar konneksies hoogty vier, en die wetenskap oor die algemeen gee al hoe meer rekenskap van die verhoudinge waarbinne 'n fenomeen ingebed is. Datawetenskap en grootdata is deel van in hierdie besef, en hierdie boek het 'n oorsig probeer verskaf oor hoe inligtingstegnologie nie alleen onderliggend is aan die hedendaagse wetenskaplike paradigma nie, maar ook hoe dit die wetenskap beïnvloed.

'n Ander tema wat 'n mens deurgaans raakloop is die betrokkenheid van die VSA se militêre intelligensie by die ontwikkeling en toepassing van inligtingstegnologie. Van die ontwikkeling van die eerste digitale rekenaars (Colossus en ENIAC), die skepping van die internet deur DARPA, die optekening van grootdata deur die NSA en CIA, netwerkontledings met behulp van i2 Analyst Notebook, Sentinel Visualizer, Palantir en Starlight VIS, tot visualiseringstegnologie soos Tableau, het militêre intelligensie altyd 'n belang in die ontwikkeling van inligtingstegnologie. Senekal (2012:473) skryf dat "militêre intelligensie aan die voorpunt van tegnologiese ontwikkeling [staan] wat betref inligtingsbestuur, en aangesien die akademiese navorser ook gekonfronteer word met die data-wolkbreuk, is die lesse wat in militêre intelligensie geleer word, bruikbaar vir die

akademie". Hierdie boek het dan ook aangetoon hoe hierdie tegnologiese hulpmiddelle binne die navorsingskonteks met vrug aangewend kan word.

Die internet beskik oor vele hulpmiddels wat die opsporing van inligting kan bespoedig, of dit nou akademiese studies, gesprekke op sosiale media of grootdatastelle is wat die navorser benodig. Die internet is 'n onuitputbare bron van inligting, en 'n goeie riglyn om te volg is om dit te benader met die wete dat as die navorser iets benodig, die kanse goed is dat dit wel deur die internet beskikbaar is – dit moet net gevind kan word. Die voorstelle wat hier gemaak is rakende soekstrategieë en webblaaie is egter geen plaasvervanger vir intuïsie nie: hoe meer vertroud die navorser met die internet word, hoe meer sal sy intuïsie hom na die regte bronmateriaal lei. Hierdie intuïsie ontwikkel egter net met ondervinding.

Rakende die ontleding van inligting; by geleentheid het 'n vriend, 'n meganiese ingenieur, aan een van die outeurs (Senekal) gesê dat hy eerder maande sal spandeer om 'n masjien te bou wat eentonige take kan outomatiseer, as wat hy daardie maande spandeer om die eentonige take self te doen. Dit is 'n deurslaggewende benaderingswyse. 'n Akademiese navorser spandeer jare om opgelei te word om sy werk te kan verrig, maar 'n onnodige groot hoeveelheid tyd word spandeer om 'donkiewerk' te verrig – die saamstel van bibliografieë en inhoudsopgawes, die liassering van bronne, ensovoorts – en in 'n grootdata-omgewing behoort dié eentonige take net toe te neem. Hierdie tyd kan meer vrugbaar spandeer word aan verdere navorsing en interpretasie, en daarom is dit sinvol om die programmatuur wat 'n navorser tot sy beskikking het eers goed onder die knie te kry. Dit is beter om 'n week te spandeer om 'n program te verken en geen daadwerklike uitkomste te kan wys nie, as wat dit is om te glo dat daar 'nie tyd' is nie, en aan te gaan sonder dat die navorser bewus is van die middele wat hy tot sy beskikking het. Hier is ook voorstelle gemaak in terme van ander programmatuur wat gebruik kan word; sommiges gratis (byvoorbeeld Qiqqa), en ander teen 'n koste (byvoorbeeld NVivo) – maar selfs wanneer rekenaarprogrammatuur kostes meebring is dit gewoonlik 'n goeie belegging.

Inligtingstegnologie verander voortdurend, en beter maniere word ontwikkel om dieselfde werk te verrig. Daarom is die belegging in infrastruktuur (in terme van tyd en geld) nie eenmalig nie; dit moet ten minste jaarliks opgedateer word. Senekal spandeer byvoorbeeld jaarliks duisende rande op rekenaarprogrammatuur en weke op navorsing oor nuwe tendense, terwyl Brokensha al breedvoerige navorsing onderneem het oor die gebruik van inligtingstegnologie vir veral onderwysdoeleindes. Let op die bronne in hierdie studie: daar is 'n groot aantal bronne wat in 2014 en 2013 gepubliseer

is, en 'n mens vergeet maklik dat fenomene soos Facebook – wat 'n groot rol gespeel het in die ontwikkeling van grootdatametodes – eers in 2006 gestig is. Hadoop is eers sedert 2006 deel van die grootdatawêreld, en die meeste grootdataprogrammatuur is maar oor die afgelope dekade ontwikkel. Die belegging in infrastruktuur werp wel vrugte af; deur bronmateriaal vinniger te kan opspoor, kan meer gelees word, wat beteken dat meer tyd beskikbaar is om nuwe terreine te ontdek en teoretiese raamwerke te ontgin, en deur nuwe rekenaarprogrammatuur te leer ken kan die navorser op 'n nuwe manier na sy onderwerp kyk.

Veral Pirolli en Card (1999) se siening van die navorser as 'n 'inligtingsroofdier' is 'n bruikbare manier om na navorsing te kyk binne die huidige universitêre opset waar begrotings en tyd al hoe meer beperk word. Meer uitsette kan met 'n kleiner tydinset gelewer word indien inligtingstegnologie doeltreffend aangewend word. Dit is ons hoop dat hierdie boek die verdere aanwending van inligtingstegnologie sal aanwakker sodat navorsingsuitsette van Suid-Afrikaanse universiteite beide kwalitatief as kwantitatief sal verhoog, en ook dat die gebruik van inligtingstegnologie tot nuwe insigte binne die Suid-Afrikaanse akademie sal lei. Watts (2011:266) se optimistiese woorde oor die gebruik van inligtingstegnologie binne die geesteswetenskappe is veral beduidend aangesien hy in fisika opgelei is, en 'n sinvolle manier om hier af te sluit:

> [N]et soos die uitvinding van die teleskoop 'n rewolusie in die bestudering van die hemel teweeggebring het, so ook deur die onmeetbare meetbaar te maak, beskik die tegnologiese revolusie in mobiele-, web- en internetkommunikasie oor die potensiaal om ons begrip van onsself en ons interaksies te verander. Merton was reg; sosiale wetenskap het steeds nie sy Kepler gevind nie. Maar drie honderd jaar na Alexander Pope aangevoer het dat die studie van die mensdom nie in die hemel nie, maar in onsself moet wees, het ons uiteindelik ons teleskoop gevind. Laat die revolusie begin ...[47]

47 Outeurs se vertaling vanuit die oorspronklike Engels.

Bibliografie

Abedin, B., F. Daneshgar & J. D'Ambra. 2014. Pattern of nontask interactions in asynchronous computersupported collaborative learning courses. *Interactive Learning Environments,* 22(1):834.

Abreu, A. & A. Acker. 2013. Context and collection: A research agenda for small data. *iConference 2013 proceedings.* 549-554.

Aghakhani, S., K. Dawoud, R. Alhajj & J. Rokne 2011. A global measure for estimating the degree of organization and effectiveness of individual actors with application to terrorist networks. In U.K. Wiil (red). *Counterterrorism and open source intelligence. New York: Springer. 189-222.*

Agrawal, D., P. Bernstein, E. Bertino, S. Davidson & U. Dayal. 2011. Challenges and opportunities with big data. Cyber Center Technical Reports, 1-15.

Alberts, D.S. & D.S. Papp (reds). 2001. *Information Age anthology: The Information Age military*. Washington: Command and Control Research Program.

Alexander, M. C. & J. A. Danowski. 1990. Analysis of ancient networks: Personal communications and the study of social structure in a past society. *Social Networks,* 12:313-335.

Allison, S., R. Heuser, M. Jockers, F. Moretti & M. Witmore. 2012. Quantitative formalism: An experiment. *N+1,* 13:81-108.

Amancio, D.R., O.N. Oliveira & L. da F. Costa. 2012. Identification of literary movements using complex networks to represent texts. *New Journal of Physics,* 14(4):043029.

Amaral, L.A.N. & J.M. Ottino. 2004. Complex networks. Augmenting the framework for the study of complex systems. *European Physical Journal,* 38:147-162.

Amaral, L.A.N., A. Scala, M. Barthélémy & H.E. Stanley. 2000. Classes of small-world networks. *PNAS,* 97:11149-11152.

Anderson, C. 2008. The end of theory: The data deluge makes the scientific method obsolete. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (22 September 2014 geraadpleeg).

Andrejevic, M. & K. Gates. 2014. Big data surveillance. *Surveillance & Society,* 12(2):185-196.

Appel, E.J. 2011. *Internet searches for vetting, investigations, and open-source intelligence*. Danvers: CRC Press.

Athenikos, S.J. 2009. Interactive visualization and exploration of information on philosophers (and artists, scholars & scientists) in an e-learning portal for digital humanities. *Symposium on Interactive Visual Information Collections and Activity (IVICA)*. 19 Junie, Austin.

Baayen, H. A. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Badge, J.L., N.F.W. Saunders & A.J. Cann. 2012. Beyond marks: New tools to visualise student engagement via social networks. *Research in Learning Technology,* 20:1-14.

Barabási, A.-L. 2003. *Linked*. London: Plume.

Barabási, A.-L. 2005a. Network theory: The emergence of the creative enterprise. *Science,* 308(5722):639-641.

Barabási, A.-L. 2005b. The origin of bursts and heavy tails in human dynamics. *Nature,* 435(7039):207-211.

Barabási, A.-L. 2009. Scale-free networks: A decade and beyond. *Science,* 325(5939):412-413.

Barabási, A.-L. 2011. *Bursts*. London: Plume.

Barabási, A.-L. 2011. The network takeover. *Nature Physics,* 8(1):14-16.

Barabási, A.-L. & R. Albert. 1999. Emergence of scaling in random networks. *Science,* 286:509-511.

Baran, P. 1964. *On distributed communications: IX. Security, secrecy, and tamper-free considerations*. Santa Monica: RAND.

Barnes, T.J. & M.W. Wilson. 2014. Big data, social physics, and spatial analysis: The early years. *Big Data & Society,* 1-14.

Bar-Yam, Y. 1997. *Dynamics of complex systems*. Colorado: Westview Press.

Bastian, M., S. Heymann & M. Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks. *Proceedings of the Third International ICWSM Conference.* 361-362.

Bavelas, A. 1948. A mathematical model for group structure. *Applied Anthropology,* 7:16-30.

Bawden, D. & L. Robinson. 2009. The dark side of information: Overload anxiety and other paradoxes and pathologies. *Journal of Information Science,* 35:180-191.

Bazeley, P. & K. Jackson. 2013. *Qualitative data analysis with NVivo*. London: Sage.

Bazzell, M. 2013. *Open source intelligence techniques: Resources for searching and analysing online information*. St. Louis: CCI.

Beckner, C., R. Blythe, J. Bybee, M.H. Christiansen, W. Croft, N.C. Ellis, J.Holland, J. Ke, D. Larsen-Freeman & T. Schoenemann. 2009. Language is a complex adaptive system: Position paper. *Language learning,* 59(1):1-26.

Beekhuyzen, J., S. Nielsen & L. v. Heller. 2010. The NVivo looking glass: Seeing the data through the analysis. *QualIT Conference – Qualitative Research in IT & IT in Qualitative Research*. November 2930, Queensland University of Technology, Brisbane.

Besser, H. 2004. The past, present, and future of digital libraries. In S. Schreibman, R. Siemens & J. Unsworth (reds). *A companion to digital humanities.* Oxford: Blackwell.

Bingham, A. 2010. The digitization of newspaper archives: Opportunities and challenges for historians. *Twentieth Century British History,* 21(2):225-231.

Blair, A. 2003. Reading strategies for coping with information overload ca. 1550-1700. *Journal of the History of Ideas,* 64(1):11-28.

Blei, D. M., T.L. Griffiths, M.I. Jordan & J.B. Tenenbaum 2004. Hierarchical topic models and the nested Chinese restaurant process. In S. Thrun, K. Saul & K. Schölkopf (reds). *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press. 17-24.

Blei, D.M., Y. Ng & M.I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research,* 3:993-1022.

Bode, K. 2012. *Reading by numbers: Recalibrating the literary field*. London: Anthem Press.

Boeker, M., W. Vach & E. Motschall. 2013. Google Scholar as replacement for systematic literature searches: Good relative recall and precision are not enough. *BMC Medical Research Methodology,* 12(131):1-12.

Boissevain, J. 1979. Network analysis: A reappraisal. *Current Anthropology,* 20(2):392-394.

Bolander, B. & M.A. Locher. 2014. Doing sociolinguistic research in computer-mediated data: A review of four methodological issues. *Discourse, Context and Media,* 3:14-26.

Bollier, D. 2010. *The promise and peril of big data*. Washington: Aspen Institute.

Bong, S. A. 2007. Debunking myths in CAQDAS use and coding in qualitative data analysis: Experiences with and reflections on grounded theory methodology. *Historical Social Research Supplement,* 19:258-275.

Borgatti, S.P., A. Mehra, D.J. Brass & G. Labianca. 2009. Network analysis in the social sciences. *Science,* 323:892-895.

Borgman, C.L. 2009. The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly,* 3(4):1-30.

Bose, R. 2008. Competitive intelligence process and tools for intelligence analysis. *Industrial Management & Data Systems,* 108(4):510-528.

Boyd, D. & K. Crawford. 2012. Critical questions for big data. *Information, Communication & Society,* 15(5):662-679.

Brannan, D. & T. Bleistein. 2012. Novice ESOL teachers' perceptions of social support networks. *TESOL Quarterly,* 46(3):519-541.

Briggs, S. 2014. Big data in education: Big potential or big mistake? http://www.innovationexcellence.com/blog/2014/01/29/big-data-in-education-big-potential-or-big-mistake (22 September 2014 geraadpleeg).

Bringer, J.D., L.D. Johnston & C.D. Brackenridge. 2004. Maximising transparency in a doctoral thesis: The complexities of writing about the use of QSR*NVivo within a grounded theory study. *Qualitative Research,* 4(2):247-265.

Broekaert, W. 2013. Financial experts in a spider web: A social network analysis of the archives of Caecilius Iucundus and the Sulpicii. *Klio: Beiträge zur Alten Geschichte,* 95(2):471-510.

Brokensha, S.I. 2012. Academic writing in Blackboard: A computer-mediated discourse analytic perspective. *Acta Academica,* 44(4):81-105.

Brokensha, S.I. & W.K. Greyling. 2014. Dispelling e-myths and pre-empting disappointment: Exploring incongruities between instructors' intentions and reality in asynchronous online discussions. *South African Journal of Higher Education.* Te perse.

Bruneel, S., K.D. Wit, J.C. Verhoeven & J. Elen. 2013. Facebook: When education meets privacy. *Interdisciplinary Journal of E-Learning and Learning Objects,* 9:125-148.

Buchanan, M. 2003. *Nexus: Small worlds and the ground-breaking science of networks*. New York: W.W. Norton & Co.

Butcher, M. 2011. Here's the guy who unwittingly live-tweeted the raid on Bin Laden. http://techcrunch.com/2011/05/02/heres-the-guy-who-unwittingly-live-tweeted-the-raid-on-bin-laden-2/ (23 Julie 2013 geraadpleeg).

Bu, D., Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li & R. Chen. 2003. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Research,* 31(9):2443-2450.

Byrne, D. & G. Callaghan. 2014. *Complexity theory and the social sciences: The state of the art*. Oxon: Routledge.

Caldarelli, G. 2013. *Scale free networks: Complex webs in nature and technology*. Oxford: Oxford University Press.

Çankaya, S., G. Durak & E. Yünkül. 2013. Using educational social networking sites in higher education: Edmodo through the lenses of undergraduate students. *European Journal of Educational Technology,* 1(1):3-23.

Castiel, L.D. & J. Sanz-Valero. 2007. Between fetishism and survival : Is the scientific article an academic commodity? *Cadernos de Saúde Pública,* 23(12):3041-3050.

Chan, R.C.H., S.K.W. Chu, C.W.Y. Lee, B.K.T. Chan & C.K. Leung. 2013. Knowledge management using social media: A comparative study between blogs and Facebook. *Proceedings of the American Society for Information Science and Technology,* 50(1):1-9.

Charmaz, K. 2014. *Constructing grounded theory*. London: Sage.

Chen, T.M. & S. Abu-Nimeh. 2011. Lessons from Stuxnet. *Security,* 91-93.

Chen, L. & T.L. Chen. 2012. Use of Twitter for formative evaluation: Reflections on trainer and trainees' experiences. *British Journal of Educational Technology,* 43(2):E49-E52.

Chen, P.D., A.D. Lambert & K.R. Guidry. 2010. Engaging online learners: The impact of web-based learning technology on college student engagement. *Computers & Education,* 54:1222-1232.

Chen, M., S. Mao & Y. Liu. 2014. Big data: A survey. *Mobile Network Applications,* 19:171-209.

Chen, M., S. Mao, Y. Zhang & V.C. Leung. 2014. *Big data-related technologies, challenges and future prospects*. Heidelberg: Springer.

Chen, C.L.P. & C.-Y. Zhang. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences,* 275:314-347.

Cherven, K. 2013. *Network graph analysis and visualization with Gephi*. Birmingham: Packt.

Choi, H. & M. Kang. 2010. Applying an activity system to online collaborative group work analysis. *British Journal of Educational Technology,* 41(5):776-795.

Christakis, N.A. & J.H. Fowler. 2007. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine,* 357(4):370-379.

Christakis, N. & J. Fowler. 2010. *Connected*. London: Harper.

Chu, S.K.W., F. Siu, M. Liang, C.M. Capio & W.W. Wu. 2013. Users' experiences and perceptions on using two wiki platforms for collaborative learning and knowledge management. *Online Information Review,* 37(2):304-325.

Clare, C. 2012. CAQDAS: Deconstructing critiques, reconstructing expectations. *MMU 15th Annual Doctoral Symposium.* Maart 14-15, *Manchester Metropolitan University*.

Clark, W., N. Couldry, R. MacDonald & H.C. Stephansen. 2014. Digital platforms and narrative exchange: Hidden constraints, emerging agency. *New Media & Society,* 1-20.

Cline, D.H. 2012. Six degrees of Alexander: Social network analysis and ancient history. *Ancient History Bulletin,* 26:1-2.

CNN Money. 2014. Fortune 500. http://money.cnn.com/magazines/fortune/fortune500/ (3 September 2014 geraadpleeg).

Costa, L. da F., O.N. Oliveira, G. Travieso, F.A. Rodrigues, P.R. Villas Boas, L. Antiqueira, M.P. Viana & L.E. Correa Rocha. 2011. Analyzing and modeling real-world phenomena with complex networks: A survey of applications. *Advances in Physics,* 60(3):329- 412.

Craig, T. & M.E. Ludloff. 2011. *Privacy and big data*. Cambridge: O'Reilly.

Crook, C. 2012. The 'digital native' in context: Tensions associated with importing web 2.0 practices into the school setting. *Oxford Review of Education,* 38(1):63-80.

Crossley, M. 2014. Global league tables, big data and the international transfer of educational research modalities. *Comparative Education,* 50(1):15-26.

Csermely, P. 2006. *Weak links: Stabilizers of complex systems from proteins to social networks*. Heidelberg: Springer.

Darvill, D. 2011. Visual analytics: Visually exploring masses of data. *Newsletter of the Association of Canadian Ergonomists*. 5-7.

Davenport, T.H. 2014. *Big data @ work: Dispelling the myths, uncovering the opportunities*. Boston: Harvard Business Review Press.

De Beer, C.S. 2003. Scholarly work: Demands, challenges and excitements. *Mousaion,* 21(1):117-136.

De Nooy, W. 1991. Social networks and classification in literature. *Poetics,* 20:507-537.

De Nooy, W. 1993. *Richtingen & lichtingen: Literaire classificaties, netwerken, instituties*. Ongepubliseerde PhD-proefskrif. Tilburg: Universiteit van Tilburg.

De Nooy, W. 2003. Fields and networks: Correspondence analysis and social network analysis in the framework of field theory. *Poetics,* 31:305-327.

Deakin, H., K. Wakefield & S. Gregorius. 2012. An exploration of peer-to-peer teaching and learning at postgraduate level: The experience of two student-led NVivo workshops. *Journal of Geography in Higher Education,* 36(4):603-612.

Defense Advanced Research Projects Agency. 2005. *Bridging the gap powered by ideas*. Arlington.

DeLaurentis, D. 2007. Role of humans in complexity of a system-of-systems. In V.G. Duffy (red). *Digital human modelling. Proceedings of the First International Conference on Digital Human Modeling, ICDHM 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007.* Berlyn: Springer. 363-371.

Deng, L. & N.J. Tavares. 2013. From Moodle to Facebook: Exploring students' motivation and experiences in online communities. *Computers & Education,* 68:167-176.

Dolowitz, D., S. Buckler & F. Sweeney. 2008. *Researching online*. London: Palgrave Macmillan.

Donnelly, D.F. & S. Boniface. 2013. Consuming and creating: Early-adopting science teachers' perceptions and use of a wiki to support professional development. *Computers & Education,* 68:9-20.

Donnelly, R. & J. Gardner. 2011. Content analysis of computer conferencing transcripts. *Interactive Learning Environments,* 19(4):303-315.

Dorogovtsev, S.N. & J.F.F. Mendes. 2001. Language as an evolving word web. *Proceedings of the Royal Society – Series B: Biological Sciences*, 268(1485):2603-2606.

Du Toit, P. & W. Smith-Muller. 2003. *Stylboek: Riglyne vir paslik skryf*. Pretoria: Van Schaik.

Dumbill, E. 2013. Making sense of big data. *Big Data,* 1(1):1-2.

Dyson, G. 2012. The dawn of computing. *Nature,* 482:459-460.

Eades, P. 1984. A heuristic for graph drawing. *Congressus Numerantium,* 42:149-160.

Faltesek, D. 2013. Big argumentation? *tripleC,* 11(2):402-411.

Fan, J., F. Han & H. Liu. 2014. Challenges of big data analysis. *National Science Review,* 1(2):293-314.

Ferguson, R. 2010. Peer interaction: The experience of distance students at university level. *Journal of Computer Assisted Learning,* 26(6):574-584.

Ferrer, C. 2013. Canonical values vs. the law of large numbers: The Canadian literary canon in the age of big data. *Rupkatha Journal on Interdisciplinary Studies in Humanities,* 5(3):81-90.

Ferrer i Cancho, R. & R.V. Solé. 2001. The small world of human language. *Proceedings of the Royal Society – Series B: Biological Sciences,* 268:2261-2265.

Fleischmann, K. 2014. Collaboration through Flickr & Skype: Can web 2.0 technology substitute the traditional design studio in higher design education? *Contemporary Educational Technology,* 5(1):39-52.

Fletcher, R. 2013. *Science, ideology, and the media: The Cyril Burt scandal*. New Brunswick, NJ: Transaction Publishers.

Fox, P. & J. Hendler. 2011. Changing the equation on scientific data visualization. *Science,* 331:705-708.

Franks, B. 2012. *Taming the big data tidal wave. Finding opportunities in huge data streams with advanced analytics*. New Jersey: John Wiley & Sons.

Freeman, L.C. 2004. *The development of social network analysis: A study in the sociology of science*. Vacouver: Empirical Press.

Friedman, H. 2003. Methodoltry and graphicacy. *American Psychologist,* 58:817-818.

Frischer, B. 2009. Art and science in the age of digital reproduction: From mimetic representation to interactive virtual reality. In A. Grande León, V. M. López-Menchero Bendicho & A. Hernández-Barahona Palma (reds). *Arqueológica 2.0. Proceedings of the 1st International Meeting on Graphic Archaeology and Informatics, Cultural Heritage and Innovation.* Sevilla: Sevilla-La Rinconada. 35-48.

Fruchterman, T.M.J. & E.M. Reingold. 1991. Graph drawing by force-directed placement. *Software: Practice and experience,* 21(11):1129-1164.

Gaertler, M. & D. Wagner 2007. Visualizing large complex networks. In G. Caldarelli & A. Vespignani (reds). *Large scale structure and dynamics of complex networks: from information technology to finance and natural science.* London: World Scientific. 115-132.

Galitski, Timothy. 2012. Reductionism gives way to systems biology. *Genetic Engineering & Biotechnology News*, 32(6):52-53.

Gallego-Arrufat, M. J., E. Gutiérrez-Santiuste & R. L. Campaña-Jiménez. 2013. Online distributed leadership: A content analysis of interaction and teacher reflections on computer-supported learning. *Technology, Pedagogy and Education,* 1-19.

Geertz, C. 1973. Thick description: Toward an interpretive theory of culture. in *The interpretation of cultures*, deur C. Geertz (red.). New York: Basic Books. 3-30.

Gehanno, J.-F., L. Rollin & S. Darmoni. 2013. Is the coverage of Google Scholar enough to be used alone for systematic reviews? *BMC Medical Informatics and Decision Making,* 13(7):1-5.

Geng, G. & L. Disney. 2014. Exploring pre-service teachers' knowledge of and ability to use text messaging. *Australian Journal of Teacher Education,* 173-182.

George, G., M.R. Haas & A. Pentland. 2014. Big data and management. *Academy of Management Journal,* 57(2):321-326.

Gibbs, G. R. 2014. Using software in qualitative data analysis. In U. Flick (red). *The Sage Handbook of Qualitative Data Analysis*. Los Angeles: Sage. 277-294.

Gilbert, L.S. 2002. Going the distance: 'Closeness' in qualitative data analysis software. *International Journal of Social Research Methodology,* 5(3):215-228.

Goh, K-I & A-L Barabási. 2008. Burstiness and memory in complex systems. *Europhysics Letters,* 81(4):480021-480026.

Gottschall, J. 2008. *Literature, science, and a new humanities*. New York: Palgrave Macmillan.

Grainger, T. & T. Potter. 2014. *Solr in Action*. Shelter Island, NY: Manning Publications.

Granovetter, M.S. 1973. The strength of weak ties. *American Journal of Sociology,* 78(6): 1360-1380.

Gries, S. 2009. *Quantitative corpus linguistics with R: A practical introduction*. New York: Routledge.

Griffiths, T.L. & M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Science* 101:5228-5235.

Guillaume, J.-L. & M. Latapy. 2004. Bipartite structure of all complex networks. *Information Processing Letters,* 90:215-221.

Guillaume, J.-L. & M. Latapy. 2006. Bipartite graphs as models of complex networks. *Physica A,* 371:795-813.

Hajič, J. 2004. Linguistics meet exact sciences. In S. Schreibman, R. Siemens & J. Unsworth (reds). *A companion to digital humanities.* Oxford: Blackwell.

Hare, J. 2014. Bring it on, big data: Beyond the hype. *Big Data,* 2(2):73-75.

Haythomthwaite, C. 1996. Social network analysis: An approach and technique for the study of information exchange. *LISR,* 18:323-342.

Hendler, J. 2013. Broad data: Exploring the emerging web of data. *Big Data,* 1(1):18-20.

Henke, G.A. 2009. *How terrorist groups survive: A dynamic network analysis approach to the resilience of terrorist organizations*. Fort Leavenworth: School of Advanced Military Studies.

Hermann, D. 2013. *Regstellende trane: Waarom verteenwoordigendheid nie gelykheid is nie*. Centurion: Kraal Uitgewers.

Hewege, C.R. & L.C.R. Perera. 2013. Pedagogical significance of wikis: Towards gaining effective learning outcomes. *Journal of International Education in Business,* 6(1):51-70.

Heymann, S. & B. Le Grand. 2013. Visual analysis of complex networks for business intelligence with Gephi. Proceedings of the *1st International Symposium on Visualisation and Business Intelligence, in conjunction with the 17th International Conference Information Visualisation. 1-6.*

Hickson, H. 2012. Reflective practice online: Exploring the ways social workers used an online blog for reflection. *Journal of Technology in Human Services,* 30(1):32-48.

Hillen, S.A. 2014. The role of discussion boards in e-collaborative learning environments (CSCL) – What kind of support can they provide?: A conceptual discussion and a qualitative case study. *Nordic Journal of Digital Literacy,* 2:128-146.

Hitzler, P. & K. Janowicz. 2013. Linked data, big data, and the 4th paradigm. *Semantic Web,* 4:233-235.

Hockey, S. 2004. The history of humanities computing. In S. Schreibman, R. Siemens & J. Unsworth (reds). *A companion to digital humanities*. Oxford: Blackwell.

Holland, J.H. 2006. Studying complex adaptive systems. *Journal of Systems Science and Complexity,* 19(1):1-8.

Honavar, V.G. 2014. The promise and potential of big data: A case for discovery informatics. *Review of Policy Research,* 31(4):326-330.

Howard, K.E., M.S. Curwen, N.R. Howard & A. Colon-Muniz. 2014. Attitudes toward using social networking sites in educational settings with underperforming Latino youth: A mixed methods study. *Urban Education,* 1-30.

Hu, Y. 2011. Algorithms for visualizing large networks. *Combinatorial Scientific Computing,* 5(3):180-186.

Humble, A.M. 2012. Qualitative data analysis software: A call for understanding, detail, intentionality, and thoughtfulness. *Journal of Family Theory & Review,* 4:122-137.

Hungerford-Kresser, H., J.L. Wiggins & C. Amaro-Jimenez. 2014. Blogging with pre-service teachers as action research: When data deserve a second glance. *Educational Action Research, 22(3):325-339*.

Ingersoll, G. 2012. Large scale search, discovery and analytics with Hadoop, Mahout and Sol. *Berlin Buzzwords*, June(4-5):1-17.

Jessop, M. 2008. Digital visualization as a scholarly activity. *Literary and Linguistic Computing,* 23(3):281-293.

Jiang, Y. & E. Katsamakas. 2010. Impact of e-book technology: Ownership and market asymmetries in digital transformation. *Electronic Commerce Research and Applications,* 9:386-399.

Jockers, M.L. 2013. *Macroanalysis: Digital methods & literary history*. Urbana: University of Illinois.

Jockers, M.L. 2014. *Text analysis with R for students of literature*. Heidelberg: Springer.

Johnson, N. 2009. *Simply complexity*. London: OneWorld.

Johnston, L. 2006. Software and method: Reflections on teaching and using QSR NVivo in doctoral research. *International Journal of Social Research Methodology,* 9(5):379-391.

Jones, B.D. & C. Breunig. 2007. Noah and Joseph Effects in government budgets: Analyzing long-term memory. *Policy Studies Journal,* 35(3):329-348.

Jones, M. & K. Diment 2010. The CAQDA paradox: A divergence between research method and analytical tool. *The International Workshop on Computer Aided Qualitative Research Asia*. The Netherlands: Merlien Institute. 82-86.

JSTOR. 2013. JSTOR Evidence in United States vs. Aaron Swartz. http://docs.jstor.org/summary.html (3 Februarie 2013 geraadpeeg).

Junior, E.V.B., A.S. Gomes & F.V. Souza. 2014. Mediating social network education teaching OOP. *American Journal of Educational Research,* 2(4):204-207.

Kamada, T. & S. Kawai. 1989. An algorithm for drawing general undirected graphs. *Information Processing Letters,* 31:7-15.

Kambatla, K., G. Kollias, V. Kumar & A. Grama. 2014. Trends in big data analytics. *Journal of Parallel Distributed Computing,* 74:2561-2573.

Keim, D., H. Qu & K.-L. Ma. 2013. Big data visualization. *IEEE Computer Graphics and Applications*, Julie/Augustus:50-51.

Kilcullen, D. 2010. *Counterinsurgency*. London: C. Hurst & Co.

Kilcullen, D. 2013. *Out of the mountains: The coming age of the urban guerrilla*. London: C. Hurst & Co.

Kinash, S., J. Brand & T.T. Mathew. 2012. Challenging mobile learning discourse through research: Student perceptions of Blackboard Mobile Learn and iPads. *Australasian Journal of Educational Technology,* 28(4):639-655.

Kirschenbaum, M.G. 2007. *The remaking of reading: Data mining and the digital humanities*. Toespraak gelewer tydens die National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation (10-12 Oktober 2007), Baltimore.

Kitchin, R. 2014. Big data, new epistemologies and paradigm shifts. *Big Data & Society,* 1(1):1-12.

Kobourov, S.G. 2013. Force-directed drawing algorithms. In R. Tamassia (red). *Handbook of graph drawing and visualization*. Danvers: CRC Press. 383-408.

Koschade, S.A. 2007. *The internal dynamics of terrorist cells: a social network analysis of terrorist cells in an Australian context*. Ongepubliseerde PhD-proefskrif. Queensland: Queensland University of Technology.

Krebs, V.E. 2002. Mapping networks of terrorist cells. *Connections,* 24(3):43-52.

Krishnan, K. 2013. *Data warehousing in the age of big data*. Amsterdam: Morgan Kaufmann.

Kroeze, J.H. 2010. The mutualistic relationship between information systems and the humanities. In K.S. Soliman (red). *Knowledge management and innovation: A business competitive edge perspective*. Cairo: IBIMA. 915-927.

Kroeze, J.H., M.C. Matthee & T.J.D. Bothma 2013. Computational information systems: Biblical Hebrew. In G. Kahn (red). *Encyclopedia of Hebrew Language and Linguistics*. Leiden: Koninklijke Brill.

Kwapień, J. & S. Drożdż. 2012. Physical approach to complex systems. *Physics Reports,* 515(3):115-226.

Landert, D. & A.H. Jucker. 2011. Private and public in mass media communication: From letters to the editor to online commentaries. *Journal of Pragmatics,* 43(5):1422-1434.

Landsberger, H.A. 1958. *Hawthorne revisited: Management and the worker – its critics, and developments in human relations in industry*. Ithaka: Cornell University.

Laney, D. 2001. 3D-data management: Controlling data – Volume, velocity and variety. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (8 Augustus 2014 geraadpleeg).

Lansing, J S. 2003. Complex adaptive systems. *Annual review of anthropology,* 32:183-204.

Latapy, M., C. Magnien & N. Del Vecchio. 2008. Basic notions for the analysis of large two-mode networks. *Social Networks,* 30:31-48.

Lazar, N. 2012. The big picture: Big data hits the big time. *Chance,* 25(3):47-49.

Leech, N.L. 2010. Interviews with the early developers of mixed methods research. In A. Tashakkori & C. Teddlie (reds). *Sage handbook of mixed methods in social & behavioural research*. Thousand Oaks: Sage. 253-272.

Lewin, K. 1951. *Field theory in social science*. New York: Harper.

Lichtman, M. 2013. *Qualitative research for the social sciences*. Los Angeles: Sage.

Lima, M. 2011. *Visual complexity: Mapping patterns of information*. New York: Princeton Architectural Press.

Loader, B.D. & W.H. Dutton. 2012. A decade in internet time. *Information, Communication & Society*, 15(5):609-615.

Long, J.C., F.C. Cunningham, J. Wiley, P. Carswell & J. Braithwaite. 2013. Leadership in complex networks: The importance of network position and strategic action in a translational cancer research network. *Implementation Science,* 8(122):1-11.

Loukides, M. 2010. *What is data science?* O'Reilly Radar.

Lues, L. & L. Lategan. 2006. *RE:search ABC*. Stellenbosch: SUN PReSS.

Luke, D.A. & K.A. Stamatakis. 2012. Systems science methods in public health: Dynamics, networks, and agents. *Annual Review of Public Health,* 33:357-376.

Lynch, C.A. 2008. The institutional challenges of cyberinfrastructure and e-Research. *EDUCAUSE Review,* 43(6):74-78.

Lyon, D. 2014. Surveillance, Snowden, and big data: Capacities, consequences, critique. *Big Data & Society,* 1(2):1-13.

MacMillan, K. & T. Koenig. 2004. The wow factor preconceptions and expectations for data analysis software in qualitative research. *Social Science Computer Review,* 22(2):179-186.

Madden, S. 2012. From databases to big data. *IEEE Internet Computing*, Junie:4-6.

Mahrt, M. & M. Scharkow. 2013. The value of big data in digital media research. *Journal of Broadcasting & Electronic Media,* 57(1):20-33.

Malkin, I. 2011. *A small Greek world: Networks in the ancient Mediterranean*. Oxford: Oxford University Press.

Malkin, I., C. Constantakopoulou & K. Panagopoulou (reds). 2011. *Greek and Roman networks in the Mediterranean*. London: Routledge.

Mandelbrot, B.B. & J.R. Wallis. 1968. Noah, Joseph and operational hydrology. *Water Resources Research,* 4:909-918.

Manovich, L. 2012. Trending: The promises and the challenges of big social data. In M.K. Gold (red). *Debates in the digital humanities.* Minneapolis: University of Minnesota Press.

Marshall, J. & H.L. Friedman. 2012. Human versus computer-aided qualitative data analysis ratings: Spiritual content in dream reports and diary entries. *The Humanistic Psychologist,* 40(4):329-342.

Martin, S., W.M. Brown, R. Klavans & K.W. Boyack. 2011. OpenOrd: An open-source toolbox for large graph layout. *SPIE Proceedings Volume 7868 Visualization and Data Analysis* 1-11.

Masucci, A.P. & G.J. Rodgers. 2006. Network properties of written human language. *Physical Review E,* 74(2):026102.

Matheson, J.L. 2005. Computer-aided qualitative data analysis software: General issues for family therapy researchers. In D.H. Sprenkle & F. P. Piercy (reds). *Research methods in family therapy.* New York: Guildford Press. 119-135.

Mattmann, C.A. & J.L. Zitting. 2011. *Tika in action*. Shelter Island, NY: Manning Publications.

Mayer-Schönberger, V. & Kenneth C. 2013. *Big data: A revolution that will transform how we live, work and think*. London: John Murray.

McAfee, A. & E. Brynjolfsson. 2012. Big data: The management revolution. *Harvard Business Review,* Oktober, 61-68.

McCallum, A.K. 2002. MALLET: A machine learning for language toolkit. mallet.cs.umass.edu (3 September 2014 geraadpleeg).

McCubbin, M. 2012. The Aftermath of Aftermath: The impact of digital music distribution on the recording industry. *University of New Hampshire Law Review,* 10(2):323-343.

McGuire, M., L. Stilborne, M. McAdams & L. Hyatt. 2000. *The internet handbook for writers, researchers and journalists*. New York: Guilford Press.

McKee, S., L. Koltutsky & M. Vaska. 2009. Introducing RefAware: A unique current awareness product. *Library Hi Tech News,* 26(9):1-6.

McNeely, C.L. & J. Hahm. 2014. The Big (Data) Bang: Policy, prospects, and challenges. *Review of Policy Research,* 31(4):304-310.

Menard-Warwick, J., A. Heredia-Herrera & D.S. Palmer. 2013. Local and global identities in an EFL internet chat exchange. *The Modern Language Journal,* 97(4):965-980.

Merico, D., D. Gfeller & G.D. Bader. 2009. How to visually interpret biological data using networks. *Nature biotechnology,* 27(10):921-924.

Meyers, W., S. Bennett & P. Lysaght 2004. Asynchronous communication: Strategies for equitable e-learning. In R. Atkinson, C. McBeath, D. Jonas-Dwyer & R. Phillips (reds). *Beyond the Comfort Zone: Proceedings of the 21st ASCILITE Conference*, Perth. 655-662.

Milgram, S. 1967. The small world problem. *Psychology Today,* 2:60-67.

Moisil, I. 2009. *Advanced methods for text retrieval*. Proceedings of the 8th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases.

Moreno, J.L. 1934. *Who shall survive?* Washington, DC: Nervous and Mental Disease Publishing Company.

Moretti, F. 2005. *Graphs, maps, trees: Abstract models for literary history*. London: Verso.

Moretti, F. 2011. Network theory, plot analysis. *New Left Review,* 68.

Motter, A.E., A.P. de Moura, Y.-C. Lai & P. Dasgupta. 2002. Topology of the conceptual network of language. *Physical Review E,* 65(6):065102.

Mouton, J. & H.C. Marais. 1990. *Basiese begrippe: Metodologie van die geesteswetenskappe*. Pretoria: RGN-Uitgewers.

Mwalongo, A. I. 2013. Peer feedback: Its quality and students' perceptions as a peer learning tool in asynchronous discussion forums. *International Journal of Evaluation and Research in Education,* 2(2):69-77.

Nadel, S.F. 1957. *The theory of social structure*. Glencoe: Free Press.

Nathans, L. & C. Revelle. 2013. An analysis of cultural diversity and recurring themes in preservice teachers' online discussions of Epstein's six types of parent involvement. *Teaching Education, 24(2):164-180.*

National Visualization and Analytics Center. 2005. Illuminating the path: The research and development agenda for visual analytics. Richland, WA: National Visualization and Analytics Center.

Neri, F. & M. Pettoni. 2009. Stalker, a multilingual text mining search engine for open source intelligence. *Advances in Soft Computing, 53:35-42.*

Newman, M.E.J. 2002. Assortative mixing in networks. Physics Review Letters, 89:2087011-2087014.

Newman, M.E.J. 2003. The structure and function of complex networks. *SIAM Review* 45(2):167-256.

Newman, M.E.J. 2010. *Networks*. Oxford: Oxford University Press.

Nguyen, L.V. 2011. Learners' reflections on and perceptions of computer-mediated communication in a language classroom: A Vietnamese perspective. *Australasian Journal of Educational Technology,* 27(8):1413-1436.

Nicolis, G. 1995. *Introduction to nonlinear science*. Cambridge: Cambridge University Press.

Nioche, J. 2013. Large scale crawling with Apache Nutch and Friends. Toespraak gelewer tydens die *Lucene/Solr Revolution EU kongres, Dublin, 1-43.*

Noruzi, A. 2005. Google Scholar: The new generation of citation indexes. Libri, 55:170-180.

O'Brien, O. & M. Glowatz. 2013. Utilising a social networking site as an academic tool in an academic environment: Student development from information-sharing to collaboration and innovation (ICI). *AISHE-J: The All Ireland Journal of Teaching and Learning in Higher Education,* 5(3):1371-13723.

Ohm, P. 2014. Changing the rules: General principles for data use and analysis. In J. Lane, V. Stodden, S. Bender & H. Nissenbaum (reds). *Privacy, big data, and the public good: Frameworks for engagement.* VSA: Cambridge University Press. 96-111.

O'Keeffe, G.S. & K. Clarke-Pearson. 2011. The impact of social media on children, adolescents, and families. *Pediatrics,* 127(4):800-804.

Olcott, A. 2012. *Open source intelligence in a networked world*. London: Continuum.

Oliveira, J.G. & A.-L. Barabási. 2005. Human dynamics: Darwin and Einstein correspondence patterns. *Nature,* 437(7063):1251.

Omand, D., J. Bartlett & C. Miller. 2012. Introducing social media intelligence (SOCMINT). *Intelligence and National Security,* 27(6):801-823.

Owen, S., R. Anil, T. Dunning & E. Friedman. 2012. *Mahout in Action*. Shelter Island, NY: Manning Publications.

Ozkan, B. C. 2004. Using NVivo to analyze qualitative classroom data in constructivist learning environments. *The Qualitative Report,* 9(4):589-603.

Padgett, J.F. & C.K. Ansell. 1993. Robust action and the rise of the Medici, 1400-1434. *The American Journal of Sociology,* 98(6):1259-1319.

Pae, J.K. 2012. How do Korean EFL students perform in computer-supported collaborative writing? *Multimedia-Assisted Language Learning,* 15(4):153-174.

Page, S.E. 2011. *Diversity and complexity*. Princeton: Princeton University Press.

Papp, D.S. & D. Alberts 1997. Preface: Technology and change in human affairs. In D.S. Papp & D. Alberts. *The Information Age: An anthology on its impact and consequences*. Washington: Command and Control Research Program. ii-viii.

Park, G.-M., S.-H. Kim, H.-R. Hwang & H.-G. Cho. 2013. Complex system analysis of social networks extracted from literary fictions. *International Journal of Machine Learning and Computing,* 3(1):107-111.

Park, H.W. & L. Leydesdorff. 2013. Decomposing social and semantic networks in emerging big data research. *Journal of Informetrics,* 7:756-765.

Patterson, E.S., D.D. Woods, D. Tinapple, E.M. Roth, J.M. Finley & G.G. Kuperman. 2001. *Aiding the intelligence analyst in situations of data overload: From problem definition to design concept exploration*. Ohio State University: Institute for Ergonomics/ Cognitive Systems Engineering Laboratory Report.

Paulus, T.M., J.N. Lester & V.G. Britt. 2013. Constructing hopes and fears around technology: A discourse analysis of introductory qualitative research texts. *Qualitative Inquiry,* 19(9):639-651.

Paulus, T.M. & G. Phipps. 2008. Approaches to case analyses in synchronous and asynchronous environments. *Journal of Computer-Mediated Communication,* 13(2):459-484.

Peitz, M. & P. Waelbroeck. 2006. Piracy of digital products: A critical review of the theoretical literature. *Information Economics and Policy,* 18:449-476.

Department of the Army and Department of the Navy. 2006. The U.S. Army and Marine Corps Counterinsurgency Field Manual. Washington.

Petty, N.J., O.P. Thomson & G. Stew. 2012. Ready for a paradigm shift? Part 2: Introducing qualitative research methodologies and methods. *Manual Therapy,* 17(5):378-384.

Pimmer, C., P. Brysiewicz, S. Linxen, F. Walters, J. Chipps & U. Gröhbiel. 2014. Informal mobile learning in nurse education and practice in remote areas: A case study from rural South Africa. *Nurse Education Today,* 34(11):1398-404.

Pirolli, P. 2007. *Information foraging theory: Adaptive interaction with information*. New York: Oxford University Press.

Pirolli, P. & S.K. Card. 1999. *Information foraging*. UIR Technical Report.

Pirolli, P. & S.K. Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *Proceedings of International Conference on Intelligence Analysis. McLean: Office of the Assistant Director of Central Intelligence for Analysis and Production. 1-6.*

Plsek, P. 2001. Redesigning health care with insights from the science of complex adaptive systems. In Committee on Quality of Health Care in America. *Crossing the quality chasm: A new health system for the 21st century*. Washington: National Academy Press. 309-322.

Polit, D.F. & C.T. Beck. 2010. Generalization in quantitative and qualitative research: Myths and strategies. *International Journal of Nursing Studies,* 47(11):1451-1458.

Polonetsky, J. & O. Tene. 2014. The ethics of student privacy: Building trust for Ed Tech. *The Digital Future of Education,* 21:25-34.

Provost, F. & T. Fawcett. 2013. Data science and its relationship to big data and data-driven decision making. *Big Data,* 1(1):51-59.

Puschmann, C. & J. Burgess. 2014. Metaphors of big data. *International Journal of Communication,* 8:1690-1709.

Quinn, N.R. & M.A. Breuer. 1979. A force directed component placement procedure for printed circuit boards. *IEEE Transactions on Circuits and Systems,* 26(6):377-388.

Rademaker, L.L., E.J. Grace & S.K. Curda. 2012. Using computer-assisted qualitative data analysis software (CAQDAS) to re-examine traditionally analyzed data: Expanding our understanding of the data and of ourselves as scholars. *The Qualitative Report,* 17(43):1-11.

Raubenheimer, J. 2012. *Doing your dissertation with Microsoft Word. A comprehensive guide to using Microsoft Word for academic writing*. Bloemfontein: True Insight.

Reichertz, J. 2004. Abduction, deduction and induction in qualitative research. In U. Flick, I. Steinke & E. von Kardoff (reds). *A companion to qualitative research*. London: Sage.

Ressler, S. 2006. Social network analysis as an approach to combat terrorism: Past, present, and future research. *Homeland Security Affairs,* 2(2):1-10.

Richards, T. 2002. An intellectual history of NUD*IST and NVivo. *International Journal of Social Research Methodology,* 5(3):199-214.

Richards, L. 2009. *Handling qualitative data: A practical guide*. London: Sage.

Ripple, A.S. 2006. Expert Googling: Best practices and advanced strategies for using Google in health sciences libraries. *Medical Reference Services Quarterly,* 25(2):97-107.

Roberts, N.C. 2011. Tracking and disrupting dark networks: Challenges of data collection and analysis. *Information Systems Frontiers,* 13:5-19.

Rodriguez, J.A. 2005. *The March 11th terrorist network: In its weakness lies its strength*. Los Angeles: XXV International Sunbelt Conference.

Rommel, T. 2004. Literary studies. In S. Schreibman, R. Siemens & J. Unsworth (reds). *A companion to digital humanities*. Oxford: Blackwell.

Rousseau, R. 2012. A view on big data and its relation to informetrics. *Chinese Journal of Library and Information Science, 5(3):12-26.*

Russom, P. 2011. Big data analytics. TDWI Research.

Ryan, M. 2009. Making visible the coding process: Using qualitative software in a post-structural study. *Issues in Educational Research,* 19(2):142-159.

Sabanci, A. & M.U. Urhan. 2014. Profiles of secondary school students' use of social media and their views about its outcomes to learning. *International Journal of Academic Research in Progressive Education and Development,* 3(1):271-284.

Said, M.N.H.M., M. Forret & C. Eames. 2014. Analysis of contradictions in online collaborative learning using activity theory as analytical framework. *Jurnal Teknologi,* 68(2):57-63.

Salganik, M.J. & D.J. Watts. 2008. Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social Psychology Quarterly,* 71(4):338-355.

Schneider, M. & M. Somers. 2006. Organizations as complex adaptive systems: Implications of complexity theory for leadership research. *The Leadership Quarterly,* 17(4):351-365.

Schoenborn, P., O. Poverjuc, V. Campbell-Barr & F. Dalton. 2013. Challenges of 'students as producers' in web 2.0: A reflective account. *Journal of Teaching and Learning with Technology,* 2(2):5-20.

Schöf, C. 2013. Big? Smart? Clean? Messy? Data in the humanities. *Journal of Digital Humanities,* 2(3):2-13.

Schreibman, S., R. Siemens & J. Unsworth 2004. The digital humanities and humanities computing: An introduction. In Schreibman, S., R. Siemens & J. Unsworth (reds). *A companion to digital humanities*. Oxford: Blackwell.

Schwarte, A., C. Haccius, S. Steenbuck & S. Steudter 2010. Usability enhancement by mining, processing and visualizing data from the Federal German Archive. *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Lisbon: Portugal. 9-14.

Scott, J. 1996. A toolkit for social network analysis. *Acta Sociologica,* 39(2):211-216.

Seeman, M. 1959. On the meaning of alienation. *American Sociological Review,* 24(6):783-791.

Senekal, J.H. 1987. *Literatuuropvattings: 'Wese' en 'waarhede' van 'n nuwe literêre teorie*. Bloemfontein: Universiteit van die Oranje-Vrystaat.

Senekal, B.A. 2011. Die digitalisering van NALN se knipselversameling: Die bemiddeling van 21ste-eeuse navorsing in die Afrikaanse letterkunde. *LitNet Akademies,* 8(2):46-65.

Senekal, B.A. 2012a. 'n Inligtingstegnologie-gesentreerde gebruikerskoppelvlak vir navorsingsdoeleindes binne die geesteswetenskappe met spesifieke verwysing na die Afrikaanse letterkunde. *LitNet Akademies,* 9(2):468-499.

Senekal, B.A. 2012b. Die Afrikaanse literêre sisteem: 'n Eksperimentele benadering met behulp van sosiale-netwerk-analise (SNA). *LitNet Akademies,* 9(3):614-638.

Senekal, B.A. 2013. 'n Netwerkontleding van die Afrikaanse poësienetwerk vanaf 2000 tot 2012. *Stilet,* 25(2):99-124.

Senekal, B.A. 2013. 'n Netwerkontleding van karakterverhoudinge in Etienne van Heerden se *Toorberg. Literator,* 34(2):1-9.

Senekal, B.A. 2013. Die gebruik van die netwerkteorie binne 'n sisteemteoretiese benadering tot die Afrikaanse letterkunde: 'n Teorie-oorsig. *Tydskrif vir Geesteswetenskappe,* 53(4):668-682.

Senekal, B.A. 2014a. An investigation of Pierre de Wet's role in the Afrikaans film industry using social network analysis (SNA). *Literator, 35(1)*.

Senekal, B.A. 2014b. *Canons and connections. A network theory approach to the study of literary systems with specific reference to Afrikaans poetry*. Washington: New Academia.

Senekal, B.A. 2014c. Dark networks: An analysis of the right wing Vaaldam plot network. *Journal for Contemporary History, 39(1):95-114*.

Senekal, B.A. 2014d. 'n Verwysingsanalise van akademiese artikels binne die Afrikaanse letterkunde. *LitNet Akademies,* 11(2):597619

Senekal, J.H. & E. Engelbrecht. 1984. *Bronne by die studie van Afrikaanse prosawerke 19001978*. Johannesburg: Perskor.

Senekal, B.A. & J.-A. Stemmet. 2014. The gods must be connected: An investigation of Jamie Uys's connections in the Afrikaans film industry using social network analysis (SNA). *Communicatio,* 40(1):1-19.

Senekal, B.A. & K. Stemmet. 2014. The South African banking director network: An investigation into interlocking directorships using social network analysis (SNA). *International Business & Economics Research Journal, 13(5):963-980*.

Senekal, J.H. & K. van Aswegen. 1980. *Bronne by die studie van Afrikaanse dramas 1900-1978*. Johannesburg: Perskor.

Senekal, J.H. & K. van Aswegen. 1981. *Bronne by die studie van die Afrikaanse digbundels 1900-1978*. Johannesburg: Perskor.

Séror, J. 2005. Computers and qualitative data analysis: Paper, pens, and highlighters vs. Screen, mouse, and keyboard. *TESOL Quarterly,* 39(2):321-328.

Shiri, A. 2014. Linked data meets big data: A knowledge organization systems perspective. *Advances In Classification Research Online,* 24(1):16-20.

Shroff, G. 2013. *The intelligent web: Search, smart algorithms, and big data*. Oxford: Oxford University Press.

Sidhu, R.K. & M.A. Embi. 2010. Learner e-tivities: Exploring Malaysian learners' roles in asynchronous computer-mediated communication. *European Journal of Educational Studies,* 2(2):157-174.

Silverman, D. 2013. *Doing qualitative research: A practical handbook*. London: Sage.

Simon, H.A. 1971. Designing organizations for an information-rich world. In M. Greenberger (red_. *Computers, communication, and the public interest.* Baltimore: Johns Hopkins Press.

Sinkovics, R.R. & E.A. Alfoldi 2012. Facilitating the interaction between theory and data in qualitative research using CAQDAS. In G. Symon & C. Cassell (reds).*Qualitative organizational research: Core methods and current challenges*. London: Sage. 109-131.

Smith, J.Z. 2012. Innovator spotlight: James Jardine and Qiqqa, a company created in light of personal need. http://www.oxbridgebiotech.com/review/careers-2/innovator-spotlight-james-jardine-and-qiqqa-a-company-created-in-light-of-personal-need (5 Junie 2013 geraadpleeg).

Smith, K., H. Brighton & S. Kirby. 2003. Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in Complex Systems,* 6(4):537-558.

Snelson, C. 2013. Vlogging about school on YouTube: An exploratory study. *New Media & Society,* 1-19.

Solé, R.V., B. Corominas-Murtra, S. Valverde & L. Steels. 2010. Language networks: Their structure, function, and evolution. *Complexity,* 15(6):20-26.

Solove, D.J. 2004. *The digital person: Technology and privacy in the information age*. New York: New York University Press.

Stewart, M.A. 2014. Social networking, workplace, and entertainment literacies: The out-of-school literate lives of newcomer adolescent immigrants. *Literacy Research and Instruction,* 53(4):347-371.

Stiller, J. & M. Hudson. 2005. Weak links and scene cliques within the small world of Shakespeare. *Journal of Cultural and Evolutionary Psychology,* 3:57-73.

Stiller, J., D. Nettle & R.I.M. Dunbar. 2003. The small world of Shakespeare's plays. *Human Nature,* 14:397-408.

Strauss, A. 1987. *Qualitative analysis for social scientists*. Cambridge: Cambridge University Press.

Strogatz, S. 2004. *Sync: The emerging science of spontaneous order*. London: Penguin.

Suderman, M. & M. Hallett. 2007. Tools for visually exploring biological networks. *Bioinformatics,* 23(20):2651-2659.

Syed, A.R., K. Gillela & C. Venugopal. 2013. The future revolution on big data. *International Journal of Advanced Research in Computer and Communication Engineering,* 2(6):2446-2451.

Szeto, E. & A.Y. Cheng. 2014. Towards a framework of interactions in a blended synchronous learning environment: What effects are there on students' social presence experience? *Interactive Learning Environments,* 1-17.

Thomas, W.G. 2004. Computing and the historical imagination. In S. Schreibman, R. Siemens & J. Unsworth (reds). *A companion to digital humanities*. Oxford: Blackwell.

Tichy, N.M., M.L. Tushman & C. Fombrun. 1979. Social network analysis for organizations. *The Academy of Management Review,* 4(4):507-519.

Tinati, R., S. Halford, L. Carr & C. Pope. 2014. Big data: Methodological challenges and approaches for Sociological Analysis. *Sociology,* 48(4):663-681.

Tracy, S.J. 2010. Qualitative quality: Eight 'big-tent' criteria for excellent qualitative research. *Qualitative Inquiry,* 16(10):837-851.

Turing, A. 1936. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society,* 2(42):230-265.

Uzum, B. 2010. An investigation of alignment in CMC from a sociocognitive perspective. *CALICO Journal,* 28(1):135-155.

Van Cleemput, K. 2012. Flemish high school students' everyday use of communication technologies for schoolwork-related communication. *Journal of Children and Media,* 6(3):367-383.

Van Dijck, J. 2014. Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & Society,* 12(2):197-208.

Van Steen, M. 2010. *Graph theory and complex networks*. Amsterdam:Van Steen.

Veletsianos, G., R. Kimmons & K. D. French. 2013. Instructor experiences with a social networking site in a higher education setting: Expectations, frustrations, appropriation, and compartmentalization. *Educational Technology Research and Development,* 61(2):255-278.

Vermeulen, W, L.O.K. Lategan & R. Litheko 2011. *The research process.* Bloemfontein: SUN PReSS.

Vicarelli, C., L. De Benedictis, S. Nenci, G. Santoni & L. Tajoli. 2013. Network analysis of world trade using the BACI-CEPII dataset. CEPII Working Paper, Augustus, 1-60.

Viljoen, H. 1986. *Die Suid-Afrikaanse romansisteem: 'n Vergelykende studie*. Ongepubliseerde PhD-proefskrif. Potchefstroom: Potchefstroomse Universiteitskollege vir Christelike Hoër Onderwys.

Von Bertalanffy, L. 1950. An outline of general system theory. *The British Journal for the Philosophy of Science,* 1(2):134-165.

Von Bertalanffy, L. 1968. *General systems theory: Foundations, development, applications*. New York: George Braziller.

Von Bertalanffy, L. 1972. The history and status of general systems theory. *The Academy of Management Journal,* 15(4):407-426.

Vorster, C. 2003. *General systems theory and psychotherapy: Beyond post-modernism*. Riviera: Satori.

Waterston, R. 2011. Interaction in online interprofessional education case discussions. *Journal of Interprofessional Care,* 25(4):272-279.

Watts, D.J. 2004. *Six degrees: The science of a connected age*. London: Vintage.

Watts, D.J. 2011. *Everything is obvious: Once you know the answer*. London: Atlantic.

Watts, D.J. & S. Hasker. 2006. Marketing in an unpredictable world. *Harvard Business Review,* 84(9):25-30.

Watts, D.J. & S.H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature,* 393(6684):409-410.

Webster, F. 1997. What information society? In D.S. Papp & D. Alberts (reds). *The Information Age: An anthology on its impact and consequences*. Washington: Command and Control Research Program.

Wiedemann, G. 2013. Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research,* 14(2).

Wiil, U.K., J. Gniadek & N. Memon 2011. A novel method to analyze the importance of links in terrorist networks. In U. K. Wiil (red). *Counterterrorism and open source intelligence*. New York: Springer. 171-188.

Wilden, A. 1980. *System and structure: Essays in communication and exchange*. New York: Tavistock.

Williams, L. & M. Lahman. 2011. Online discussion, student engagement, and critical thinking. *Journal of Political Science Education,* 7(2):143-162.

Wooldridge, R. 2004. Lexicography. In S. Schreibman, R. Siemens & J. Unsworth (reds). *A companion to digital humanities*. Oxford: Blackwell.

Young, N.S., J.P.A. Ioannidis & O. Al-Ubaydli. 2008. Why current publication practices may distort science. The market for exchange of scientific information: The winner's curse, artificial scarcity, and uncertainty in biomedical publication. *PLOS Medicine,* Oktober, 1-19.

Zimmer, M. 2010. But the data is already public: On the ethics of research in Facebook. *Ethics and Information Technology,* 12(4):313-325.

Zumel, N. & J. Mount. 2014. *Practical data science with R*. Shelter Island, NY: Manning Publications.

Indeks

Hierdie boek is toegespits op navorsers en doen verslag oor navorsing wat oor die afgelope paar jaar onderneem is om vas te stel hoe inligtingstegnologie aangewend is en kan word vir navorsingsdoeleindes binne die geesteswetenskappe, sowel as watter implikasies die gebruik van inligtingstegnologie vir die geesteswetenskappe inhou in die inligtingsera. Die beginsels, implikasies, probleme en geleenthede van inligtingstegnologie en die digitale revolusie word teen die agtergrond van groot-data bespreek, en word veral in verband gebring met die geesteswetenskappe in Suid-Afrika.

Surfers van die Tsunami besin oor die groot verskuiwing wat in die afgelope paar dekades plaasgevind het in die wyse waarop inligting versamel, ontleed, aangebied en versprei word. Terwyl die nuwe tegnologie reeds in die natuurwetenskappe wyd gebruik word, staan die geesteswetenskappe volgens die outeurs nou eers aan die begin van 'n groot omwenteling ...

Die boek bied 'n diepgaande oorsig van internasionale kennis en literatuur – metodes en voorbeelde van teksontginning word deeglik bespreek.

Prof JH Kroeze
Universiteit van Suid-Afrika (Unisa)

Burgert A Senekal is sedert 2008 verbonde aan die Universiteit van die Vrystaat, en is tans 'n post-doktorale navorsingsgenoot by die Eenheid vir Taalfasilitering en Bemagtiging. Sy onlangse navorsingsbelangstelling sluit sisteemteorie, netwerkteorie, en inligtingstegnologie in, veral waar inligtingstegnologie ingespan kan word om binne die netwerkteorie komplekse sosiale sisteme te ontleed.

Susan Brokensha is 'n senior lektor in die Departement Engels aan die Universiteit van die Vrystaat en het 'n PhD in Toegepaste Linguistiek. Haar navorsing fokus op rekenaar-bemiddelde kommunikasie, veral die studie van diskoers gegenereer via sinchrone en asinchrone vorme van kommunikasie, maar ook die diskoers op sosiale netwerk webwerwe.

sb SUNBONANI SCHOLAR

9 781920 382643 >