Die digitalisering van NALN se knipselversameling: Die bemiddeling van 21ste-eeuse navorsing in die Afrikaanse letterkunde

Burgert A. Senekal
Departement Afrikaans en Nederlands, Duits en Frans
Universiteit van die Vrystaat

Summary

The digitisation of NALN's collection of newspaper clippings: Enabling 21st-century research in Afrikaans literature

The contemporary world is an extremely complex environment due to globalisation and the internet. Within this globalised interdependent framework, researchers, both in an academic context and in non-academic settings such as business, cannot expect to succeed without incorporating resources that extend the reach of the individual's environment and expedite the processing of information. Not only has the amount of information circling the globe increased rapidly over the past two decades, but museums and libraries have had to downsize their staff, which effectively means that fewer resources are available to handle more information and in addition serve a larger population. If the humanities' basic tasks are "preserving, reconstructing, transmitting, and interpreting the human record" (Frischer 2009:15), technology is the key, enabling research across international borders, distributing data, findings and insights globally, and managing the deluge of information that is characteristic of the 21st-century world. Within this context, digitisation – the process of converting analogue documents to a digital format – occupies a privileged position, enabling the distribution of information globally (and thereby contributing to information overload), as well as safeguarding the preservation of important historical material. Digitisation has become a global trend, and although many South African heritage institutions have been slow to make the transition from analogue to digital – mainly because of budget constraints – museums and archives in South Africa have realised the potential that digitisation holds, and are now digitising their collections.

Digitisation has numerous benefits, including making information available globally and adding to a museum's archival role by providing faithful copies to researchers without exposing the original documents to the elements and to wear and tear, and because handling digital material saves time, enabling staff to attend to the collection, retrieval, and preservation of analogue documents. In addition, digitisation enables using information technology for research purposes, including in lexicography, linguistics, historiography, and in literary analysis. This allows the analysis of data sets that are much larger than a human researcher can handle, and abroad researchers in what is known as the digital humanities recently began campaigning for the better integration of information technology in social science research. Schwarte et al. (2010:9) argue, "The study of history usually requires the extraction of relevant data out of large corpora. This data extraction is typically very time-consuming if done manually; time which is afterwards lacking for the actual research. It is therefore urgently necessary to make data on the one hand accessible, and on the other hand

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

automate the process of data extraction and presentation." However, in order for advanced software to utilise these large sets of information a significant body of documents has to be available in a suitable digital format.

After obtaining funding from the private sector in late 2009 the digitisation project conducted by the University of the Free State's Department of Afrikaans and Dutch, German and French at the National Afrikaans Literary Museum and Research Centre (NALN) in Bloemfontein was started early in 2010. The project involves digitising newspaper clippings and magazine and journal articles numbering over 300 000 documents, covering literary reviews, polemics and other Afrikaans cultural matters. As such, the so-called clippings collection is unique and comprehensive, and researchers working on Afrikaans literature have used this collection extensively. Borgman (2010:10) specifically mentions newspaper articles as one of the types of data necessary for research in the humanities, together with letters, books, articles, diaries and more. The digitisation of such a collectionordinarily has the twin goals of preservation and access: "Firstly, by making a digital copy of old and fragile materials, the handling of the original document is reduced. Second, having made a digital copy, the document may be accessed simultaneously by many more users in diverse locations, subject only to whatever access controls are appropriate for the document, while the original may be accessed only from within the Special Collections area" (Thomas, Cramond, Emery and Scott 2005:10). The primary objective of the current digitisation project is to enhance NALN's preservation function, but with Afrikaners now dispersed across the globe, digitisation also enables the more efficient distribution of information to researchers overseas, since NALN is both the guardian of Afrikaans's literary heritage and a research centre. Executed by the author under the guidance of Prof. H.P. van Coller, the project has only recently found its feet, but in addition to the benefits this project holds for the museum, the project can also contribute to establishing Afrikaans literature more firmly within the global information network, and creates new possibilities for research in Afrikaans literature, both qualitatively and quantitatively.

This article examines the historical background of the project, the progress made so far, and practical lessons learned – from scanning resolutions and file formats to working within budget constraints and with a limited number of personnel. Throughout the discussion of scanning, image enhancement, conversion to Portable Document Format (PDF), Optical Character Recognition (OCR), metadata and Information Retrieval (IR) the article examines different available options, the shortcomings and benefits associated with each, and why this project chose to take a specific route. Because numerous similar projects have been undertaken overseas, the article shows how those involved in this project have learnt from publications on other digitisation projects, adapting lessons learnt to the constraints and requirements of the project after extensive research and testing. The article furthermore notes that digitisation requires a sharp, never-ending learning curve: cyber-infrastructure is not adequate, because it is currently sufficient, but must stay updated and keep pace with international developments. In following publications in the digital humanities, the article lastly explores opportunities created by digitisation to conduct research in a new way through advanced data mining and visualisation software. It is argued that this has the potential to stimulate a fresh appreciation of the printed text: "The development of digital collections does not require the destruction of books [or other sources]; instead, it may provoke more interest in their existence and provide different opportunities for their study through keyword and structured searching" (Willett 2004).

This digitisation project is neither the largest of its kind in the country nor the first, but has already managed to contribute to the debate on the digitisation of heritage material in South Africa – a contribution this article aims to make as well.

Keywords: *Digitisation, digital humanities, information technology, archives, NALN*

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

Opsomming

Digitalisering het vele voordele, insluitend die wêreldwye beskikbaarstelling van inligting, die aanvulling van ’n museum se bergingsfunksie deur getroue kopieë aan navorsers te verskaf sonder dat die oorspronklike dokumente aan die elemente en slytasie onderwerp word, en tydbesparings vir personeel wat hulle in staat stel om meer aandag aan die versameling, ontsluiting en bewaring van dokumente te bestee. Verder bemiddel digitalisering die gebruik van inligtingstegnologie vir navorsingsdoeleindes, soos reeds in die buiteland onder andere in die leksikografie, linguistiek, historiografie, en heelwat in die letterkunde onderneem is. Sodoende kan die ontleding van datastelle onderneem word wat veel groter is as wat ’n menslike navorser kan hanteer, en in die buiteland het navorsers binne wat as die digitale geesteswetenskappe bekend staan, hulself onlangs begin beywer om inligtingstegnologie beter binne geesteswetenskaplike navorsing te integreer. Ten einde gevorderde sagteware te kan benut om hierdie groot hoeveelhede inligting te hanteer, moet ’n beduidende korpus analoogdokumente egter digitaal beskikbaar wees. Crane, Babeu en Bamman (2007:120) skryf: “We need to increase physical and intellectual access to every type of content and we need methods that are automated and can be applied to large bodies of content.” Die Universiteit van die Vrystaat se digitaliseringsprojek by die Nasionale Afrikaanse Letterkundige Museum en Navorsingsentrum (NALN) in Bloemfontein het pas sy voete gevind, maar bo en behalwe die voordele wat dit vir die museum inhou, kan die projek ook daartoe bydra om die Afrikaanse letterkunde stewiger in die globale inligtingsnetwerk te vestig én nuwe moontlikhede vir navorsing binne die Afrikaanse letterkunde skep. Binne die raamwerk van sowel buitelandse digitaliseringsprojekte as die digitale geesteswetenskappe bespreek hierdie artikel die vordering wat binne die eerste jaar gemaak is, die lesse wat geleer is, en die potensiaal wat digitalisering binne die Afrikaanse letterkunde inhou.

Trefwoorde: Digitalisering, digitale geesteswetenskappe, inligtingstegnologie, argiewe, NALN

1. Inleiding

Die hedendaagse wêreld is ’n uiters komplekse omgewing as gevolg van globalisasie en die internet. Binne hierdie geglobaliseerde interafhanklike konteks kan navorsers, sowel in ’n akademiese sfeer as in nie-akademiese opsette soos die sakewêreld, nie verwag om sukses te behaal sonder die inkorporering van middele wat die reikwydte van die individu se leefwêreld vergroot en die verwerking van inligting bespoedig nie. Boiangiu, Spataru, Dvornic en Bucur (2008:67) voer aan: “The future of cultural development throughout the world is demanding more than ever for a larger distribution and a better preservation in time of knowledge”, en tegnologie is die sleutel hiervoor. Sedert die vroeë steentydperk was tegnologie altyd die oplossing om met beperkte mannekrag beter resultate op te lewer; en binne enige organisasie – museum, besigheid of universiteit – verhoog tegnologie nie alleen produktiwiteit nie, maar beskik dit ook oor die vermoë om ’n kwalitatiewe verbetering van dienslewering te bewerkstellig deur die menslike dimensie meer vaartbelyn te maak. Soos Nuance Software (2010) se webblad dit stel: “Every 12 filing cabinets require an additional employee to maintain. And this is just the beginning of what it costs to manage paper in an increasingly digital world.” Ten spyte van die groei van die algemene bevolking en

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

ook van studentegetalle oor die afgelope twee dekades, het museums en biblioteke begrotingsbesnoeiings beleef: die James Hardiman Biblioteek by die Nasionale Universiteit van Ierland in Galway het byvoorbeeld ’n 11,7 persent-besnoeiing in personeel beleef (Cox 2010:7), en NALN se personeel is ook drasties besnoei en herontplooi. Die behoefte bestaan dus om met minder personeel meer te vermag, en hierdie tendens blyk wêreldwyd toe te neem.

Die hoeveelheid inligting het ook radikaal toegeneem sedert die algemene beskikbaarstelling van rekenaars en die internet in die negentigerjare. Die direkteur van Google, Eric Schmidt, beweer dat die mensdom van aanvang tot 2003 vyf eksagrepe (*exabytes*) se data gegenereer het, terwyl daar tans vyf eksagrepe elke twee dae gegenereer word (Tynan 2010). Schmidt noem dit ’n “data tsunami”: een eksagrepe bestaan uit 1018 grepe, met ander woorde een een miljoen teragrepe (Skokowski 2010). Dit kom neer op 125 miljoen geheuestokkies (*flash drives*) van vier gigagrepe elk wat per dag tot die globale inligtingsnetwerk toegevoeg word.

Crane, Babeu en Bamman (2007:117) glo akademici moet radikaal nuwe tegnologie en sosiale konvensies ontwikkel om te kan bou op die “sterrestelsels data” wat nou vorm aanneem, en beperkte vordering is reeds gemaak: universiteite het toegang tot hoëspoed-internet en aanlyn-argiewe van akademiese joernale wat navorsing bespoedig; kommunikasie geskied vinniger deur middel van e-posse; en navorsing word vinniger in ’n publiseerbare formaat verwerk deur middel van woordverwerkingsprogramme.

Die gebruik van inligtingstegnologie het sowel kwantitatiewe as kwalitatiewe implikasies: aangesien minder tyd spandeer word om deur argiewe te soek sonder dat enigiets gevind word, kan meer tyd spandeer word om die inhoud van dokumente te ontleed (Bingham 2010:229).

Lynch (2008) glo dat die impak van tegnologie op die wetenskap breed beskou moet word:

> When we speak of the changes wrought by information technology, we consider information technology in its broadest sense: not only high-performance computing and advanced computer communication networks but also sophisticated observational and experimental devices and sensor arrays attached to the network, as well as software-driven technologies such as high-performance data management, data analysis, mining and visualisation, collaboration tools and environments, and large-scale simulation and modelling systems.

Die gevorderde sagteware (programmatuur) waarna Lynch verwys, is egter steeds nie deel van die hoofstroom binne die geesteswetenskappe in die buiteland of in Suid-Afrika nie, en Borgman (2010:3) noem dat die natuurwetenskappe die geesteswetenskappe vooruit is in die VSA en die Verenigde Koninkryk (VK), waar onderskeidelik na *kuberinfrastruktuur* en *eScience* verwys word:

Digital scholarship remains emergent in the humanities, while eScience has become the norm in the sciences. The humanities need not emulate the sciences, but can learn useful lessons by studying the successes (and limitations) of cyberinfrastructure and eScience initiatives.

Afgesien van die algemene gebruik van die internet, woordverwerkers en programme soos EndNote en Mendelay vir akademiese doeleindes, beskik inligtingstegnologie oor

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

die vermoë om navorsing op ’n nuwe manier te benader. In die buiteland het die begrip *digitale geesteswetenskappe* onlangs opgang begin maak (vir ’n oorsig oor dié ontwikkeling, sien Hockey 2004:3–17). Frischer (2009:15) definieer *digital humanities* as “the application of information technology as an aid to fulfill the humanities’ basic tasks of preserving, reconstructing, transmitting, and interpreting the human record”.

Inligtingstegnologie is onder andere al aangewend in die leksikografie (Wooldridge 2004), die linguistiek (Hajič 2004), die historiografie (bv. Thomas 2004; Schwarte, Haccius, Steenbuck en Steudter 2010), en die letterkunde (bv. Rommel 2004). In alle fasette waarmee die geesteswetenskappe hulle bemoei, kan tegnologie dus aangewend word, en behoort dit gebruik te word ten einde nie binne die “data-tsoenami” verlore te raak nie.

Digitale dokumente skep die geleentheid om navorsing anders te benader, veral deur die ontleding van groter datastelle (Borgman 2010:8). Schreibman, Siemens en Unsworth (2004:xxvi) let daarop dat ’n rekenaar ’n navorser in staat stel om verbintenisse tussen tekste en terme, patrone, samevoegings en afwesighede te identifiseer wat die navorser nie daarsonder sou kon doen nie. Vir hulle bied die digitale geesteswetenskappe “a new horizon for humanities scholarship, a paradigm as powerful as any that has arisen in any humanities discipline in the past”.

Craig (2004:282–5) noem ook hoe rekenaarmatige stilistiese analise kan help om dispute oor skrywerskap van ouer tekste op te los, en die impak wat dit kan hê op ons begrip van die literêre kanon. Die volume inligting wat hiervoor ontleed moet word, is egter veels te groot vir ’n mens om akkuraat te verwerk en dus skep die rekenaar moontlikhede om tot heel nuwe insigte rakende die (Afrikaanse) literêre sisteem te kom. Deur middel van digitale koerantartikels kan ’n navorser byvoorbeeld uitvind wanneer ’n term sy verskyning gemaak het, of wanneer ’n kwessie in die media opgeduik het – so het ’n Amerikaanse navorser byvoorbeeld ’n vroeër gebruik van die woord *jazz* ontdek as wat voorheen bekend was (Bingham 2010:228). Data-ontginning is ook al gebruik om sentimentaliteit in 19de-eeuse Amerikaanse fiksie te ontleed (Horton, Taylor, Yu en Xiang 2006), asook die representasie van geslagtelikheid in Shakespeare se werke (Hota, Argamon, Koppel en Zigdon 2006).

Die digitalisering van analoogdokumente is egter die fondament waarop alle ander tegnologiese middele berus: sonder bruikbare digitale weergawes van dokumente bestaan daar geen moontlikheid om gevorderde sagteware aan te wend om groter volumes data te verwerk nie. Borgman (2010:11) is deel van ’n koor navorsers binne die digitale geesteswetenskappe wanneer hy vra dat ’n “kritiese massa” digitale hulpbronne opgebou word om meer algemene studies met ’n groter reikwydte te bemagtig. Schwarte e.a. (2010:9) voer aan:

> The study of history usually requires the extraction of relevant data out of large corpora. This data extraction is typically very time-consuming if done manually; time which is afterwards lacking for the actual research. It is therefore urgently necessary to make data on the one hand accessible, and on the other hand automate the process of data extraction and presentation.

Digitale dokumente het ook verskeie ander voordele bo analoogdokumente, en hier word slegs ’n paar genoem in navolging van Rothenberg en Bikson (1999:69):

- Digitale dokumente kan foutloos gekopieer word, wat beteken dat dit aan studente en kollegas deurgegee kan word sonder om die kwaliteit van die dokument te verminder. ’n Kenmerk van analoogkopiëring (fotostate) is dat

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

daar altyd 'n mate van ruis (*noise*) met elke kopie intree, wat beteken dat inligting verlore kan gaan.

- Digitale dokumente kan oor netwerke versprei word. In plaas daarvan dat 'n fotostaat gemaak moet word om aan 'n kollega te gee, kan dit bloot aangeheg word in 'n e-pos, wat tyd en geld bespaar.

- Digitale dokumente kan deur 'n wye verskeidenheid media versprei word, insluitend CD's en DVD's, asook geheuestokkies en inderdaad papieruitdrukke. Dit maak inligting makliker oordraagbaar en gevolglik meer vryelik beskikbaar, wat Malhan (2006:6) 'n "demokratisering" van inligting noem: "Everyone having connectivity to the Internet can access the same information at the same time using search engines. Scientists in the developing countries can access research journals at the same time their counterparts do in the developed countries."

- Die elektroniese opspoor van dokumente, asook die soek vir inligting binne dokumente, is eksponensieel vinniger. Digitale dokumente wat op 'n persoonlike rekenaar of op 'n bediener gestoor word, kan byvoorbeeld met behulp van elektroniese soekfunksies opgespoor word.

Sien ook Malhan (2006:5) in dié verband.

Digitalisering voldoen gevolglik aan twee kernvereistes vir navorsing in die geesteswetenskappe: "We need to increase physical and intellectual access to every type of content and we need methods that are automated and can be applied to large bodies of content" (Crane e.a. 2007:120).

Om hierdie redes, asook bogenoemde mannekragbesnoeiings, het die Universiteit van die Vrystaat (UV) sedert 2008 betrokke geraak by die digitalisering van NALN se versameling koerantknipsels. Hierdie artikel bepreek die vordering wat gemaak is, die lesse wat geleer is, en die potensiaal wat digitalisering binne die Afrikaanse letterkunde inhou.

2. Agtergrond

Borgman (2010:4) glo wetenskaplikes moet self betrokke raak by die implementering van inligtingstegnologie: "Only those who do the work and who require the infrastructure are in a position to take the field forward. Librarians and technology developers are essential partners, but those who conduct the research must take the lead." Ná eerste samesprekings in 2007 met betrokkenes in die Vrystaatse Provinsiale Regering het Hennie van Coller van die UV in 2008 begin met onderhandelings om befondsing te verkry vir 'n digitaliseringsprojek by NALN. Befondsing is uiteindelik in Oktober 2009 ontvang, maar as gevolg van verskeie oponthoude het die digitaliseringsprojek eers in Mei 2010 begin. Intussen is die UV se projek op administratiewe vlak by die Erfenisstigting se Korttermyningrypingsplan (KTIP)[1] ingelyf. Ten spyte daarvan dat die projek deur die privaatsektor befonds word en die UV-personeel hiervoor kontrakteer, bly dit 'n projek in diens van en tot voordeel van NALN alleenlik (behalwe dat enigiets wat tot NALN se voordeel strek, indirek ook die Afrikaanse gemeenskap bevoordeel).

NALN se knipselversameling bewaar items wat handel oor 'n geraamde 8 000 Afrikaanse skrywers, 'n versameling wat strek oor nagenoeg 125 jaar en 'n geskatte 350

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

000 tot 400 000 knipsels insluit (Liebenberg 2009:1). Party van hierdie knipsels is oorspronklikes wat op A4-papiere geplak is, terwyl ander fotostate is, en nog ander in hul oorspronklike publikasievorm is. 'n Knipsel bestaan egter gereeld uit 'n hele aantal bladsye ('n artikel uit 'n tydskrif tel byvoorbeeld as een knipsel), en daarom is die aantal bladsye wat gedigitaliseer moet word, veel meer as die aantal knipsels.

Die digitalisering van hierdie tipe versameling het primêr ten doel om inligting te bewaar, asook om toegang aan alle belanghebbende partye te verleen (Gatos, Mantzaris, Perantonis en Tsigris 2000:77). Thomas, Cramond, Emery en Scott (2005:10) skryf:

> Digitization projects are about *preservation* and *access*. Firstly, by making a digital copy of old and fragile materials, the handling of the original document is reduced. Second, having made a digital copy, the document may be accessed simultaneously by many more users in diverse locations, subject only to whatever access controls are appropriate for the document, while the original may be accessed only from within the Special Collections area (skrywers se beklemtoning).

In NALN se geval moet die knipselversameling beskikbaar wees vir navorsingsdoeleindes vanaf skoolvlak tot die vlak van professionele navorsers – en nie alleen landswyd en binnelands nie: met Afrikaanssprekendes deesdae dikwels op ander kontinente en met Afrikaans wat al hoe meer belangstelling in die buiteland wek, stel digitalisering navorsers in staat om vanaf enige plek toegang tot NALN se knipselversameling te verkry. Borgman (2010:10) noem juis koerantknipsels as een van die vorme van data wat noodsaaklik is vir navorsing binne die geesteswetenskappe, tesame met briewe, boeke, artikels, dagboeke ensovoorts. Alhoewel die meeste van hierdie ander formate voorlopig buite die bestek van die huidige digitaliseringsprojek val, word wel beoog om na afloop van die voltooiing van die knipselversameling na hierdie versamelings uit te brei.

Terwyl onderhandelings vir befondsing deur Van Coller onderneem is, het ek navorsing onderneem oor die waarde en nut van digitalisering, asook oor die werkswyse met betrekking daartoe, en die hoof van NALN, Otto Liebenberg, is ook gekontak om NALN se behoeftes te bepaal. Projekte soos die Gutenberg-projek, Towards A New Age of Partnership (TANAP), die Million Book Project (MBP), Gallicia, en American Memory, asook pogings deur maatskappye soos Google, Yahoo en Amazon, verskaf waardevolle lesse in digitalisering wat by die plaaslike projek geïntegreer is. So verskaf Rothenberg en Bikson (1999) byvoorbeeld 'n waardevolle analise van die digitalisering van die Nederlandse Nasionale Argief, en Pramod, Ambati, Pratha en Jawahar (2006) praktiese lesse in verband met die Million Book Project (MBP).

3. Toerusting

Soos Pramod e.a. (2006:426) opmerk, benodig 'n digitaliseringsprojek gepaste hardeware (apparatuur), soos skandeerders en rekenaars, en die regte sagteware om die projek binne 'n realistiese tydraamwerk te voltooi. Hier moet die projekleier hom vergewis van wat op die mark bestaan, én van wat die behoeftes van die projek is. Wat hardeware aanbetref, moet die regte skandeerder en rekenaar(s) gekies word om die formaat van die oorspronklike dokumente te akkommodeer – dit is byvoorbeeld oorbodig om 'n duurder oorhoofse skandeerder aan te skaf vir koerantknipsels wat op

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

A4-papier geplak is, maar 'n sogenaamde *flatbed* is totaal ongeskik vir die skandering van groot getalle boeke.

Sagteware verg 'n nóg groter assessering van die beskikbare middele en die vereistes van die projek: om sagteware spesifiek vir die betrokke digitaliseringsprojek te ontwikkel, is ideaal, soos in die geval van die Australiese digitalisering van koerante (Holley 2007:6–7), maar tans bestaan die kundigheid hiervoor nie binne die digitalisering van NALN se knipsels nie, en om 'n privaatmaatskappy te kontrakteer, is bloot te duur. Daar bestaan boonop reeds talle programme wat in die behoeftes van so 'n projek kan voorsien.

'n Kwessie wat tydens die digitaliseringswerksessie by die Universiteit van Pretoria (UP) op 11 en 12 November 2010 geopper is, was watter rol sogenaamde oopbron-sagteware (*open source*-sagteware) kan speel om insetkoste te verlaag. Gratis databasis-programme soos MySQL, mSQL en PostgreSQL word wêreldwyd in die digitale wetenskappe benut (Ramsay 2004:186), maar Platt (2010:4) waarsku:

> Often the question might arise: to open source, or not to open source? While using open source software is the current trend, institutions should look closely at their resources to determine if they can support the technological and human resources required to work with open source software packages.

Omdat oopbron-sagteware nie verhandel word nie, word dit nie vergesel van tegniese ondersteuning of opleiding nie. Neem ook in ag dat baie oopbron-programme 'n verdere inset van die gebruiker verg omdat dit nie volledig is nie. Platt (2010:4) skryf byvoorbeeld met betrekking tot DSpace: "In DSpace, the level of programming needed to make customizations beyond changing the color scheme of the website can be daunting for someone without experience in both programming and website design." Terwyl Platt wel die tegniese-ondersteuningstruktuur gehad het om oopbron-sagteware te gebruik, het die digitaliseringsprojek by NALN eerder gesteun op sagteware wat kommersieel beskikbaar is.

4. Die digitaliseringsproses

Digitalisering verwys na al die stappe betrokke in die omskakeling van 'n hardekopie of papierkopie (analoogkopie) na 'n elektroniese (digitale) kopie (Singh 2003:12), insluitend die toevoeging van metadata. Pramod e.a. (2006:428–30) onderskei tussen vyf fases tydens die digitaliseringsproses: skandering, verwerking, herkenning en rekonstruksie, gehaltebeheer, en die oplaai van digitale dokumente op die webblad. Daar bestaan volgens Willett (2004:243–4) verskeie enkoderingsopsies:

- Eerstens kan 'n elektroniese transkripsie gemaak word, wat by uitstek die mees kompakte vorm van 'n teks is en gevolglik die minste geheue van die rekenaar en bediener opneem, wat ook beteken dat dit die maklikste oor netwerke versprei kan word, selfs wanneer die eindverbruiker 'n stadige internetkonneksie het. Die mees akkurate benadering volgens Deegan en Tanner (2004:494) is die sogenaamde hertik-tegniek (*rekeying*-tegniek): twee of drie tiksters tik die betrokke dokument oor, en dan word sagteware gebruik om die verskille tussen die drie weergawes te bepaal (Microsoft Word 2007 kan byvoorbeeld so 'n vergelyking doen). Die kanse dat drie tiksters dieselfde fout sal maak, is minimaal, maar die nadeel hiervan is dat dit tyd- en gevolglik geldrowend is, en die formaat van die dokument gaan ook verlore.

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

- Tweedens kan ’n digitale beeld geskep word in formate soos TIFF of JPEG (sien hier onder), maar dit het die nadeel dat volteksoektogte bemoeilik word.

- Derdens bestaan die opsie van ’n geënkodeerde teks, gewoonlik in Text Encoding Initiative (TEI), wat uitstekende toegang tot die inligting verleen, maar soos die eerste opsie tyd- en geldrowend is, en ook die formaat van die dokument vernietig.

- Die vierde opsie word deur die Gutenberg-projek gevolg en behels die enkodering van inligting in die eenvoudigste vorm moontlik: ’n gewone teksdokument (.txt). Dieselfde nadele van die eerste en derde opsies geld, maar projek Gutenberg se stigter, Michael Hart, glo hierdie opsie skep die grootste waarskynlikheid dat die digitale dokument steeds in die afsienbare toekoms leesbaar sal wees.

- Die vyfde opsie is om ’n soekbare PDF met behulp van karakterherkenningsagteware te skep. Alhoewel die akkuraatheid swakker is as byvoorbeeld by die eerste opsie, is die kostebesparings noemenswaardig, en soos hier onder in meer besonderhede verduidelik word, is hierdie opsie vir die digitalisering van NALN se knipselversameling gekies.

Die digitaliseringsproses by NALN verloop as volg:

4.1 Skandering

Koerantknipsels word met ’n resolusie van 300 spd (*stippels per duim* (spd))na Tagged Image File Format (TIFF) geskandeer, aangesien dié formaat ’n internasionale standaard is (soos byvoorbeeld gebruik deur die UV) én makliker bewerk word as dokumente in Portable Document Format (PDF). TIFF en PDF, tesame met Extensible Markup Language (XML), is die algemeenste formate wat wêreldwyd in digitale databasisse gebruik word (Hodge en Frangakis 2004:29; sien ook Platt 2010:7). Beide TIFF en PDF is die voorkeurformaat wanneer die oorspronklike vorm van ’n dokument belangrik is, soos die geval is met die Victorian Electronic Records Strategy (VERS) en ook NALN (XML behou nie die oorspronklike formaat nie). Tydens die digitalisering van die IIUM Biblioteek in Maleisië is daar ook na TIFF geskandeer (Abdullah en Marsidi 2008:20), asook tydens die digitalisering van Australië se koerante (Holley 2007:7). JPEG (Joint Photographic Experts Group) is ongeskik vir swart-en-wit dokumente en word vir geen digitaliseringsprojek aanbeveel nie, aangesien dit ’n vorm van kompressie is waardeur besonderhede verlore gaan. Ook kan ’n mens nie ’n samestelling in JPEG skep nie, maar wel in TIFF en PDF.

Aanvanklik sou ’n resolusie van 600 spd gebruik gewees het (in navolging van die Million Books Project), maar ’n hoër resolusie veroorsaak dat die digitale dokumente meer besonderhede bevat, wat ook beteken dat meer ruis (*noise*) ontstaan. 300 spd word in elk geval deur Nuance se Omnipage, ABBYY Finereader en SimpleSoftware se SimpleIndex aanbeveel, en tydens die digitalisering van die IIUM Biblioteek – wat ook teks eerder as foto’s gedigitaliseer het – is ’n resolusie van tussen 200 en 300 spd gebruik (Abdullah en Marsidi 2008:20). Skandering word meestal na grys gedoen, aangesien dit beter vir karakterherkenning- oftewel Optical Character Recognition (OCR)-sagteware is. Serfontein (2011) van die UV se rekenaardienste voer aan:

> Met 300 spd kry jy die beste akkuraatheid tydens OCR sonder om spoed of lêergrootte prys te gee. Indien faktore soos OCR-akkuraatheid, skanderingspoed, OCR-spoed en lêergrootte in ag geneem word, kan daar by ’n

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

lesing van 200 spd na 300 spd 'n verbeteringsgaping waargeneem word wat dubbel so groot is as enige ander metings. Van 300 spd na 400 spd is die verbeteringsgaping baie klein, maar nog steeds daar voor dit begin afneem.

'n Verdere voordeel van 'n resolusie van 300 spd eerder as 600 spd is die grootte van die geheue wat benodig word vir TIFF-dokumente: soos New Hampshire Universiteit 'n "megabyte shock" (Platt 2010:7) ervaar het toe hulle 600 spd-TIFF-skanderings 30 tot 80 megagrepe per bladsy opgeneem het, het die digitalisering van NALN se knipselversameling ook gereeld die rekenaar se geheue opgebruik. Groter dokumente neem langer om tussen rekenaars of tussen die rekenaar en bediener oor te dra, en neem meer spasie op enige bergingsmedium (CD, DVD of hardeskyf) op. Verheldering is een manier om die hoeveelheid geheue wat 'n dokument opneem, te reduseer (sien volgende onderafdeling), maar wanneer teks eerder as foto's gedigitaliseer word, is 'n resolusie van 300 spd ideaal. Wanneer tydskrifartikels met kleurfoto's egter gedigitaliseer word, word dit teen 'n resolusie van 600 spd ingeskandeer.

4.2 Verheldering

Analoogkopiëring (bv. fotostate), asook die omskakeling vanaf analoog na digitaal, veroorsaak dat 'n mate van ruis ontstaan. In die geval van die digitalisering van NALN se koerantknipsels moet daar boonop ook dokumente gedigitaliseer word wat reeds met 'n swak fotostaatmasjien gekopieer is, asook die standaardruis wat onvermydelik is. Dit is van kardinale belang dat digitale dokumente skoon is: "The key to 'better' data – that is, data suitable for curation, reuse, and sharing – is capturing data as cleanly as possible and as early as possible in its life cycle" (Borgman 2010:13). Om hierdie rede word voorkeur aan oorspronklike dokumente bo fotostate verleen, indien moontlik. Waar kommersiële karakterherkenningsagteware betrokke is, is "high accuracy [...] usually available only with some of the recognition languages, on very clean scans with little or no background images and relatively simple layouts" (Hu, Rose en Bederson 2009:396).

Met duurder hardeware word die hoeveelheid ruis beperk, maar die begroting van die huidige digitaliseringsprojek is relatief klein (die Fujitsu fi-5900c kos byvoorbeeld nagenoeg tien keer meer as ons HP n8460). Sagteware soos Accusoft se ScanFix, wat in die Million Book Project gebruik is, kan andersins ruis effektief verminder, maar dit is ook duurder as wat ons begroting toelaat.

Soektogte op die internet het ook 'n verskeidenheid belowende programme opgelewer (soos Readiris en ClearImage), maar Paperport se Auto Enhance- funksie het uiteindelik 'n gepaste oplossing gebied. Dit is 'n maklike, eenklik-manier om 'n koerantberig se agtergrond te verwyder en die teks vir karakterherkenning te verhelder, en is een van die goedkoopste opsies beskikbaar. Indien nodig word TIFF-dokumente verder in Adobe Photoshop skoongemaak, kleur-enkodering gestandaardiseer, ensovoorts.

Aangesien die projek met groot volumes werk, en beperkte mannekrag het (slegs twee personeellede), word die digitaliseringsproses sover moontlik geoutomatiseer, en prosesse is uitgewerk in bykans elke program waarmee gewerk word wat dokumente nie een vir een nie, maar in honderde of duisende verwerk, indien moontlik. Gehaltebeheer verg egter altyd 'n menslike inset.

Hier is 'n voorbeeld van 'n dokument wat sodanig verhelder is:

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

Stigting meen Afrikaans moet sy amptelike status behou

Korrespondent

BLOEMFONTEIN. – Die Stigting vir Afrikaans meen dat Afrikaans sy amptelike status in 'n nuwe politieke bedeling moet behou, het mnr. Tom de Beer, voorsitter van die stigting, gesê.

Mnr. De Beer het kommentaar gelewer op die taalbeleid wat die ANC vandeesweek op sy kultuur- en ontwikkelingskonferensie in Johannesburg aanvaar het.

Daarin word onder meer voorgestel dat gelyke status verleen word aan alle tale wat in Suid-Afrika gepraat word.

Die ANC se taalkommissie het ook bevind dat Afrikaans en Engels ander tale onderdruk het.

Mnr. De Beer gee toe dat die ander inheemse tale in die verlede afgeskeep is, en dat dit vorentoe groter erkenning moet kry. Maar om te beweer dat Afrikaans (en Engels), soos die ANC se taalkommissie aanvoer, ander tale onderdruk het, is 'n wanvoorstelling.

"Engels moet maar sy eie verdediging behartig. Afrikaans moes sedert sy ontstaan self 'n stryd om erkenning en oorlewing stry. 'n Goeie voorbeeld daarvan is die weerstand in die begin van die eeu teen pogings van die Engelse koloniale owerheid om Engels aan Afrikaanssprekendes op te dwing."

Afrikaanssprekendes het dus begrip vir die behoefte van ander moedertaalsprekers dat hul tale ook erken word, het mnr. De Beer gesê.

Die Stigting vir Afrikaans se standpunt is dan ook dat hy die ander tale niks misgun nie – hy beywer hom bloot vir Afrikaans.

Per slot van rekening, Afrikaans is lank reeds as amptelike taal gevestig en voldoen aan al die vereistes wat aan so 'n taal gestel word. Hy is volledig ontwikkel as kommunikasie-, kultuur-, wetenskap- en regstaal.

Daar sou dus met oortuiging aangevoer kan word dat Afrikaans 'n verworwe reg het om, naas Engels, as amptelike taal voort te bestaan, ten minste tot van die ander tale dieselfde ontwikkelingsvlak bereik het.

Mnr. De Beer het gesê 15,2 miljoen Suid-Afrikaners kan Afrikaans gebruik. Dit is sowat die helfte van die totale bevolking. Dit is meer as enige ander taal.

Afrikaans kan dus 'n waardevolle brug tussen die verskillende bevolkingsgroepe wees, en dit is presies waarop die Stigting hom in sy verskillende aksies toespits.

Die Stigting het byvoorbeeld reeds 'n geletterdheidsprogram ingestel wat binnekort sterk uitgebrei gaan word.

Stigting meen Afrikaans moet sy amptelike status behou

Korrespondent

BLOEMFONTEIN. – Die Stigting vir Afrikaans meen dat Afrikaans sy amptelike status in 'n nuwe politieke bedeling moet behou, het mnr. Tom de Beer, voorsitter van die stigting, gesê.

Mnr. De Beer het kommentaar gelewer op die taalbeleid wat die ANC vandeesweek op sy kultuur- en ontwikkelingskonferensie in Johannesburg aanvaar het.

Daarin word onder meer voorgestel dat gelyke status verleen word aan alle tale wat in Suid-Afrika gepraat word.

Die ANC se taalkommissie het ook bevind dat Afrikaans en Engels ander tale onderdruk het.

Mnr. De Beer gee toe dat die ander inheemse tale in die verlede afgeskeep is, en dat dit vorentoe groter erkenning moet kry. Maar om te beweer dat Afrikaans (en Engels), soos die ANC se taalkommissie aanvoer, ander tale onderdruk het, is 'n wanvoorstelling.

"Engels moet maar sy eie verdediging behartig. Afrikaans moes sedert sy ontstaan self 'n stryd om erkenning en oorlewing stry. 'n Goeie voorbeeld daarvan is die weerstand in die begin van die eeu teen pogings van die Engelse koloniale owerheid om Engels aan Afrikaanssprekendes op te dwing."

Afrikaanssprekendes het dus begrip vir die behoefte van ander moedertaalsprekers dat hul tale ook erken word, het mnr. De Beer gesê.

Die Stigting vir Afrikaans se standpunt is dan ook dat hy die ander tale niks misgun nie – hy beywer hom bloot vir Afrikaans.

Per slot van rekening, Afrikaans is lank reeds as amptelike taal gevestig en voldoen aan al die vereistes wat aan so 'n taal gestel word. Hy is volledig ontwikkel as kommunikasie-, kultuur-, wetenskap- en regstaal.

Daar sou dus met oortuiging aangevoer kan word dat Afrikaans 'n verworwe reg het om, naas Engels, as amptelike taal voort te bestaan, ten minste tot van die ander tale dieselfde ontwikkelingsvlak bereik het.

Mnr. De Beer het gesê 15,2 miljoen Suid-Afrikaners kan Afrikaans gebruik. Dit is sowat die helfte van die totale bevolking. Dit is meer as enige ander taal.

Afrikaans kan dus 'n waardevolle brug tussen die verskillende bevolkingsgroepe wees, en dit is presies waarop die Stigting hom in sy verskillende aksies toespits.

Die Stigting het byvoorbeeld reeds 'n geletterdheidsprogram ingestel wat binnekort sterk uitgebrei gaan word.

Volumes verander die werkswyse egter geheel en al. Nuance se Paperport kan byvoorbeeld nie die Auto Enhance-funksie toepas op veel meer as 500 dokumente op een slag nie, terwyl Adobe Acrobat nie veel meer as 1 000 dokumente op 'n slag omskakel na soekbare PDF nie. Ook is karakterherkenning akkurater wanneer met kleiner hoeveelhede gewerk word. Met toetse van 250 dokumente is hierdie probleme dus nie uitgewys nie, en daar moes telkens teruggegaan word om die werkswyse van voor af uit te werk. Hierdie tendens het ook 'n belangrike les opgelewer: daar moet nooit gewerk word aan die meester-TIFF-dokumente nie, sodat foute nie die herskandering van dokumente noodsaak nie. Ook beteken die berging van die TIFF-dokumente dat nóg 'n formaat van internasionale standaard gebruik word, wat die moontlikheid verhoog dat digitale dokumente in die komende dekades steeds bruikbaar sal wees.

Die totale versameling koerantknipsels vanaf 1993 tot 2001 is nou reeds geskandeer en elektronies verhelder – 74 141 digitale dokumente altesaam.

4.3 Omskakeling na PDF

NALN benodig nie bloot teks nie, maar dokumente wat, sover sinvol, getrou aan die oorspronklikes is, natuurlik met inagneming daarvan dat die digitale kopie altyd sal verskil van die analoogkopie, omdat dit binne 'n ander formaat is (Hutcheon 2006:16). Soos Borgman (2010:9) te kenne gee: "[O]ne person's noise is often another person's signal"; die formaat waarin dokumente verskyn, dra self betekenis, en daarom is teks alleen nie die enigste oorweging by digitalisering nie. Internasionale instansies soos die Victorian Electronic Records Strategy (VERS) en JSTOR (Journal Storage) verkies juis PDF-dokumente, omdat dit die oorspronklike vorm van 'n dokument behou. PDF word onder andere as formaat gebruik deur die International Union of Crystallography (IUCr) – een van die leiers op die gebied van digitalisering (Hodge en Frangakis 2004:16) – asook die VSA se National Records Administration (Hodge en Frangakis 2004:29) en die Profiles in Science Project van die Nasionale Museum van Geneeskunde in die VSA (Hodge en Frangakis 2004:39). Ook het New Hampshire Universiteit hierdie formaat vir verspreidingsdoeleindes gekies (Platt 2010:7), want "Information in the

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

repository should be not just available, but accessible to all. This includes maintaining file sizes to enable faster load times, ensuring that even users with dial-up modems can download the files in a reasonable amount of time" (Platt 2010:8).

PDF is deur Adobe ontwikkel en in 1993 geloods, en is tans 'n ISO-standaard (ISO 32000) (Adobe Systems Incorporated 2011).

Daar bestaan verskillende tipes PDF-enkoderings:

1. PDF wat nie soekbaar is nie (geen karakterherkenning-sagteware is gebruik nie, byvoorbeeld SA Media se dokumente op hul webblad).

2. PDF waar die teks onttrek is (wanneer 'n TIFF omgeskakel word na 'n teks en dan na 'n PDF omgeskakel word, soos deur FineReader of OmniPage).

3. PDF wat verskuilde teks agter die beeld enkodeer.

Tipes 1 en 3 lyk identies, maar 3 bemagtig 'n voltekssoektog, wat tipe 1 nie toelaat nie. Tipe 2 is altyd 'n skoner, meer leesbare dokument, maar verg dat 'n kontroleerder die karakterherkenning verifieer – daarsonder word skandeerfoute deur die PDF gereproduseer.

Hier is 'n voorbeeld van 'n omskakeling met behulp van FineReader:

DIE verengelsing van die sakewêreld en die gevolglike miskenning van Afrikaans in die advertensiewese is 'n ou probleem. Dit is egter nie 'n probleem wat sommer weggewens kan word nie. Afrikaanssprekendes sal dit eenvoudig nie toelaat nie.

Die feit van die saak is dat Afrikaanssprekendes feitlik alle fasette van die Suid-Afrikaanse ekonomie met hul koopkrag oorheers, soos pas weer baie duidelik uitgespel is deur 'n omvattende navorsingstuk van Nasnet, bemarkingsafdeling van Naspers se drie Afrikaanse koerante, Die Burger, Beeld en Die Volksblad.

DIE verengelsing van die sakewêreld en die gevolglike miskenning van Afrikaans in die advertensiewese is 'n ou probleem. Dit is egter nie 'n probleem wat sommer weggewens kan word nie. Afrikaanssprekendes sal dit eenvoudig nie toelaat nie.

Die feit van die saak is dat Afrikaanssprekendes feitlik alle fasette van die Suid-Afrikaanse ekonomie met hul koopkrag oorheers, soos pas weer baie duidelik uitgespel is deur 'n omvattende navorsingstuk van Nasnet, bemarkingsafdeling van Naspers se drie Afrikaanse koerante, Die Burger, Beeld en Die Volksblad.

Tipes 1 en 3 **Tipe 2**

'n Verdere nadeel van tipe 2 is dat strepe, handskrif en lettertipes verlore gaan, wat beteken dat die digitale kopie nie getrou is aan die analoogkopie nie. Om hierdie rede is besluit om na tipe 3 PDF-dokumente oor te skakel aangesien dit 'n middeweg (of wat Deegan en Tanner (2004:496) 'n "hibridiese oplossing" noem) verteenwoordig: dit bly getrou aan die oorspronklike, maar bemagtig 'n voltekssoektog.

Alhoewel tipe 3 nie so akkuraat is soos 'n tipe 2 wat deur 'n kontroleerder nagegaan is nie, wys bostaande voorbeeld van tipe 2 dat die karakterherkenningsagteware in hierdie paragraaf 100 persent akkuraat was – grotendeels te danke aan die gehalte van die TIFF. Die verlies van inligting wat met 'n tipe 3-omskakeling gepaard gaan, word dus deur effektiewe voorbereiding beperk. Deegan en Tanner (2004:496) skryf juis dat so 'n hibridiese vorm baie geskik is vir koerantknipsels, veral omdat dit toegang tot groot volumes inligting bemagtig terwyl onkoste heelwat laer is as by ander opsies. Foute kan in elk geval oor die hoof gesien word in soektogte deur van sogenaamde "fuzzy matching" gebruik te maak: 'n soekenjin (soos Paperport) verwag dan klein herkenningsfoute te wagte wanneer deur die teks gesoek word.

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

'n Sekuriteitsenkodering is noodsaaklik om die verspreiding en gesaghebbendheid van digitale dokumente te reguleer, en Adobe Acrobat word hier as een van die markleiers geag; dit is dan ook in die digitalisering van die IIUM Biblioteek aangewend (Abdullah en Marsidi 2008:19). NALN se digitale dokumente sal uiteindelik ook voorsien word van 'n sekuriteitsenkodering en logo wat beide NALN se eiendomsreg van die oorspronklike erken en NALN se gesaghebbendheid behou. Hier is natuurlik ook vele ander opsies: dokumente kan digitaal onderteken word om gesaghebbendheid te verseker, of ander annotasies kan outomaties oor die hele versameling ingevul word wat NALN se eiendomsreg identifiseer; ensovoorts. Soos altyd word opsies vir hulle goedkeuring aan NALN voorgelê.

4.4 Karakterherkenning

Sagteware soos Nuance OmniPage, ABBYY FineReader en Adobe Acrobat word algemeen gereken as die markleiers op die gebied van karakterherkenningsagteware (Abdullah en Marsidi 2008:15), en die Million Book Project (MBP) het 90 tot 95 persent akkuraatheid met behulp van ABBYY FineReader verkry (Pramod e.a. 2006:434). Hierdie program is ook aangewend by die Universiteit van New Hampshire (Platt 2010:8), en tydens die digitalisering van Australië se koerante (Holley 2007:13). Kae en Learned-Miller (2009:5) noem egter dat OmniPage 15 (die nuwe weergawe is 17) in hulle eksperimente die akkuraatste program was, alhoewel dit later oortref is deur die oopbron-program Tesseract.

Alhoewel FineReader dus die bekendste is, verskil akkuraatheid van projek tot projek, en die verskille tussen die genoemde drie programme se akkuraatheid en aankooppryse is so minimaal dat die keuse van 'n program afhang van die gebruiker se persoonlike voorkeur.

Die Text+Berg-projek gebruik beide FineReader en Omnipage, en vergelyk dan die resultate in 'n karakterherkenningsweergawe van die hertik-metode. In hul projek was FineReader die akkuraatste (99,26 persent teenoor OmniPage se 96,21 persent), maar die kombinasie het resultate verbeter, en hulle ondersoek ook die gratis program Tesseract (Volk, Marek en Sennrich 2010:63–5). By die toekomstige digitalisering van manuskripte – wat 'n groter herkenningsakkuraatheid verg – kan Text+Berg se voorbeeld gevolg en al vier programme gebruik word.

Aangesien die digitaliseringsprojek by NALN PDF's genereer wat tot bogenoemde tipe 3-PDF hoort, word karakterherkenningsresultate nie nagegaan nie. Deegan en Tanner (2004:496) skryf dat dit ook die werkswyse is van die Forced Migration Online-projek, want "What is important to users of FMO is documents or parts of documents dealing with key topics, rather than that they can retrieve individual instances of words or phrases." Só 'n werkswyse sal egter nie deug wanneer manuskripte gedigitaliseer word vir potensiële geoutomatiseerde linguistiese of stilistiese analise – wat 'n akkuraatheid van meer as 99,99 persent verg – nie, maar aangesien die knipselversameling dieselfde funksie verrig as die dokumente van FMO, beteken die bykomende onkoste verbonde aan die regstelling van herkenningsfoute dat hierdie opsie tans te duur is.

In die toekoms kan moontlikhede ondersoek word om 'n Afrikaanse woordeboek met hierdie sagteware te integreer, wat die akkuraatheid van die karakterherkenning sal verbeter (Hu e.a. 2009:396).

Die TIFF-dokumente van alle skanderings word dus nie alleenlik behou ter wille van bewaring nie, maar ook in afwagting van toekomstige tegnologiese ontwikkeling:

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

verbeterings in karakterherkenning soos byvoorbeeld deur Kae en Learned-Miller (2009) geïllustreer, kan toekomstige resultate heelwat verbeter.

AnyDoc en Cardiff Teleform is twee seldsame programme wat handskrif kan herken, en kan in die toekoms oorweeg word om oorspronklike korrespondensie en handgeskrewe manuskripte elektronies leesbaar te maak. Beide het egter hul beperkings, en die digitalisering van handgeskrewe manuskripte sal in die toekoms veel makliker deur middel van tiksters voltrek kan word, met die oorspronklike bloot 'n gewone PDF wat nie soekbaar is nie.

4.5 Metadata

Die term *metadata* het alombekend geraak sedert 1994, en kan volgens die Association for Library Collection and Technical Services (ALCTS) se taakspan oor metadata soos volg gedefinieer word: "Metadata are structured, encoded data that describe characteristics of information bearing entities to aid in the description, discovery, assessment and management of the described entities" (Singh 2003:16). Sonder bruikbare metadata kan 'n gebruiker nie digitale dokumente vind nie, en metadata is dus 'n noodsaaklike komponent van enige digitaliseringsprojek. Dublin Core is een van die internasionale standaarde (NISO Standaard Z39.85) vir die skep van metadata (Platt 2010:6). Dit is 'n vorm van wat Besser (2004:564) ontdekkingsmetadata (*discovery metadata*) noem, maar daar bestaan ook ander vorme, soos strukturele metadata, wat gebruik word om die bladsye van 'n digitale boek om te blaai, en administratiewe metadata, wat gebruik word om seker te maak dat verskillende bladsye van 'n digitale boek bymekaar bly. Aangesien die digitaliseringsprojek by NALN verskillende bladsye reeds in die regte volgorde kombineer, word hier slegs na ontdekkingsmetadata verwys.

NALN beskik reeds oor 'n databasis (Inmagic DB/Text) waarop alle metadata (trefwoorde, publikasies, skrywers, datums, onderwerpe, ensovoorts) ingevul is. DB/Text is spesifiek vir biblioteekdoeleindes ontwikkel, en aangesien 'n groot gedeelte van NALN se versamelings uit dokumente bestaan, is dit veral hiervoor geskik. Om duplisering te vermy, koppel ons gevolglik nie metadata direk aan digitale dokumente self nie, maar skakel eerder digitale dokumente met die databasis deur middel van 'n hiperskakel (*hyperlink*) op die databasis.

Hierdie koppeling beteken dat die digitale dokument dieselfde naam moet dra as die aanwinsnommer op die databasis, en die herbenoeming van digitale dokumente in ooreenstemming met die oorspronklike knipselnommer was een van die grootste uitdagings van die projek. Aangesien die getal knipsels op tussen 350 000 en 400 000 geskat word, is die individuele herbenoeming van digitale dokumente nie 'n koste-effektiewe oplossing nie. Neem ook in ag dat knipsels gereeld uit meer as een bladsy bestaan (tussen 1 en 25 bladsye), dus is die aantal enkelbladsy-dokumente wat herbenoem moet word, veel meer as die geskatte 400 000 knipsels. Omdat 'n knipsel nie noodwendig uit een bladsy bestaan nie, is daar geen konstante volgorde waarop outomatiese herbenoemingsagteware berus nie en daarom was ons verplig om dokumente individueel te herbenoem. Adobe Bridge is vir 'n tyd lank gebruik omdat dit 'n makliker gebruikerskoppelvlak (*user interface*) verskaf as Windows Explorer, en later is 'n Acrobat-*plug-in* (Autosplit) opgespoor wat die proses met bykans 50 persent bespoedig het. Laasgenoemde opsie verg dat bladsye in 'n enkele PDF van honderde bladsye op 'n slag gekombineer word; hierna word boekmerke (*bookmarks*) ingesit wat dieselfde naam as die knipselnommer dra, en dan word die dokument weer outomaties geskei (*split*) sodat die nuwe dokument – 'n knipsel met al sy bladsye – dieselfde naam as die boekmerk (en daarom ook die oorspronklike knipselnommer) dra. Dit lyk

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

heelwat meer omslagtig as wat dit is: met hierdie metode herbenoem 'n mens dan slegs tussen 'n derde en die helfte van die dokumente. Die grootte van dokumente was egter 'n probleem, omdat 'n dokument wat uit meer as 1 000 bladsye bestaan, te lomp geword het om mee te werk.

Die grootste komponent van die probleem is dat knipselnommers in handskrif op die bladsye aangebring is. Karakterherkenningsagteware kan nie handskrif lees nie, en eksperimente is gedoen met ReadIris Pro om te kyk of dié sagteware nie ook die knipselnommer akkuraat kan lees nie. Toetse was egter onsuksesvol, en in opvolg is Perceptive Software se ImageNow en Cardiff se Teleform ondersoek, maar ook sonder sukses. SimpleSoftware se SimpleIndex verskaf egter 'n makliker gebruikerskoppelvlak wat dit moontlik maak om dokumente individueel maar vinnig te herbenoem. Ook het SimpleIndex 'n funksie om outomaties veelbladsy-dokumente te genereer, wat beteken dat knipsels wat oor meer as een bladsy strek, nie later gekombineer hoef te word nie. Om 'n tydrowende herhaling te vermy, word dokumente dus met behulp van SimpleIndex herbenoem en gekombineer voordat karakterherkenningsagteware gebruik word om die teks soekbaar te maak.

4.6 Inligtingsherwinning

Soektogte kan op NALN se databasis gedoen word, wat dan aan die digitale dokument gekoppel is. NALN se personeel kan byvoorbeeld op die databasis soek na knipsels oor Etienne Leroux, en dan word 'n lys van knipsels opgelewer wat deur middel van die trefwoorde gevind word. Hierna kan die knipsel gedruk of ge-e-pos word.

Die ideaal is dat hierdie soekfunksies ook aan die algemene verbruiker beskikbaar sal wees deur middel van die internet, soos byvoorbeeld die geval met JSTOR se webblad is. Kopieregkwessies moet egter nog uitgeklaar word; soos in die geval met die IIUM Biblioteek kan dit moontlik beteken dat slegs geregistreerde gebruikers toegang tot die volteks van dokumente kan verkry (Abdullah en Marsidi 2008:7); of die gebruiker kan 'n lys van knipsels saamstel en deur NALN se personeel aanvra.

Volteksoektogte kan ook gedoen word vanaf 'n platform soos Adobe Acrobat, Stottler Henke se Aware, of Christian Ghisler se Total Commander – of selfs Windows 2007 se soekfunksie – wat dan deur die teks van al die dokumente soek vir trefwoorde of -frases. Sodoende kan elke voorkoms van 'n term opgeroep word, alhoewel hierdie metode gewoonlik langer neem as 'n databasis-soektog wat slegs deur die metadata soek.

Nuance se PaperPort 12 Pro het onlangs belowende resultate opgelewer toe dié program se soekfunksie gebruik is om volteksoektogte uit te voer en dit binne sekondes deur 60 000 bladsye kon soek. Boonop is die toets gedoen op dokumente wat nog nie deur die karakterherkenningsproses was nie – hier is deur TIFF-beelde gesoek *sonder* die gebruik van OmniPage, FineReader of Acrobat.

Data-ontginning- of data-herwinningsagteware soos WordCruncher of Eaagle Full Text Mapper kan ook moontlik gespesialiseerde soektogte en dataverwerking bemagtig, wat nuwe moontlikhede vir navorsing in die Afrikaanse letterkunde skep. Laasgenoemde program help met die visualisering van inligting, soos Craig (2004:273–7) doen wanneer hy Shakespeare se dramateks rekenaarmatig vergelyk, en Berzak, Richter en Ehrler (2010) in die vergelyking van Fidel Castro se toesprake. Ook Steve Ramsay het reeds sodanige eksperimente onderneem met die ontleding van Shakespeare se *Anthony and Cleopatra* (Warwick 2004:375). Volgens Warwick het hierdie eksperimente letterlik 'n heel nuwe visie van die betrokke drama geskep het, en Jessop (2008:281) skryf:

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

> The numerous visual metaphors that are used to describe our cognitive processes hint at the nexus of relationships between what we see and what we think. We say we "see" when we mean that we "understand"; we try to organize and make our ideas "clear" by bringing them into "focus", and so on. When faced with tasks that require substantial thought or organisation of ideas we will often reach for a pen and paper to "sketch out" (another visual metaphor) our thoughts.

Sulke sagteware sou byvoorbeeldkon help om te bepaal watter skrywers die sigbaarste in NALN se versameling van die eerste dekade van die 21ste eeu is, asook in watter posisie hulle val én wanneer hulle die sigbaarste was. Sulke data kan gebruik word om die ontwikkeling van die Afrikaanse literêre kanon na te vors. Vir Warwick (2004:379) is dit een van die mees opwindende geleenthede wat rekenaars binne die geesteswetenskappe bied, want "A new way of looking at a text can lead to a way of reading it that is unconstrained by the bindings of the printed medium."

Beide databasis- en volteksoektogte het voor- en nadele, maar in beginsel is die vermoë om volteksoektogte te doen bloot aanvullend tot NALN se bestaande stelsel. Dit skep nuwe navorsingsmoontlikhede, en vergemaklik die opspoor van inligting, maar vervang geensins die analoogversameling of die werk wat kundiges reeds gedoen het nie.

5. Volhoubaarheid

Wanneer die UV se projek voltooi is, moet NALN uiteraard kan voortgaan met die digitalisering van knipsels soos dit daagliks tot die versameling toegevoeg word. 'n Personeellid van NALN is reeds opgelei om 2010 se knipsels te skandeer en te stoor, en sy het reeds 1 627 bladsye ingeskandeer. Soos die projek verder ontwikkel, sal NALN se personeel bemagtig word om die hele proses selfstandig te behartig.

6. Slot

Berging van inligting is slegs een funksie van inligtingsinstellings soos NALN; die ander is die verlening van toegang tot inligting en die opbou en instandhouding van versamelings (Rothenberg 1999:2). Die digitalisering van NALN se argief moet dus gesien word as een element in die museum se strategiese raamwerk, en nie as vervanging van die analoogversameling nie. Digitalisering sal vir seker die bergingsfunksie aanvul deur getroue kopieë aan navorsers te verskaf sonder dat die oorspronklike dokumente aan die elemente en slytasie onderwerp word, en omdat dit tydbesparings vir personeel sal meebring wat hulle in staat sal stel om meer aandag aan die versameling, ontsluiting en bewaring van dokumente te bestee, maar digitalisering kan nie die analoogversameling vervang nie. Soos Willett (2004) skryf:

> The development of digital collections does not require the destruction of books [of ander bronne - BS]; instead, it may provoke more interest in their existence and provide different opportunities for their study through keyword and structured searching.

As 'n digitale verwerking van 'n analoogkopie het digitalisering ook die potensiaal om 'n nuwe kyk na die analoogkopie te prikkel: "Adaptation is repetition, but repetition

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

without replication" (Hutcheon 2006:7) – nuwe inligting ontstaan altyd. Besser (2004:558) skryf ook:

> Though the promise of digital technology in almost any field has been to let one do the same things one did before but better and faster, the more fundamental result has often been the capability of doing entirely new things.

Die potensiaal van digitalisering om die navorsingsmoontlikhede binne die Afrikaanse letterkunde uit te brei was een van die eerste lesse wat in die eerste jaar van die digitaliseringsprojek geleer is, en dit berus op 'n volledige korpus digitale dokumente.

'n Aantal verdere lesse is geleer. Die belangrikste hiervan is die besef dat digitalisering 'n baie groot mate van aanpasbaarheid verg: soos die formaat van dokumente, personeelvaardighede en ander omstandighede verander, bied tegnologie altyd al hoe beter oplossings, maar 'n projekleier moet bereid wees om aanpasbaar te bly. Dit veronderstel ook dat digitalisering 'n skerp, nimmereindigende leerkurwe het: kuberinfrastruktuur is nie perfek omdat dit tans voldoende is nie, maar moet bygewerk bly en in pas met internasionale ontwikkelinge. Platt (2010:10) het tot 'n bykans identiese slotsom gekom:

> Learning and implementing standards for metadata, master files and access files was time-consuming, but taking the time to establish standards in the beginning doubtless saved a great deal of trouble for the future. Even so, it will be necessary to keep up with developing industry standards, and it would not be surprising if further adjustments are needed down the road. A digital repository is much like a physical building; periodic maintenance, remodelling, and wear and tear should be anticipated and expected. (Sien ook Holley 2007:6.)

Die digitaliseringsprojek by NALN het pas sy voete gevind, maar die vordering wat hier gemaak word, kan nie alleen NALN help om 'n hoër gehalte diens aan navorsers en die publiek te lewer nie, dit kan ook help om die Afrikaanse letterkunde stewiger in die globale inligtingsnetwerk te vestig.

Bibliografie

Abdullah, S. en S. Marsidi. 2008. Digitisation of Arabic materials in IIUM Library: Challenges and problems. *World Congress of Muslim Librarian and Information Scientists*, Putra.

Adobe Systems Incorporated 2011. *Adobe fast facts*. http://www.adobe.com/aboutadobe/pressroom/pdfs/fastfacts.pdf (1 Februarie 2011 geraadpleeg).

Berzak, Y., M. Richter en C. Ehrler. 2010. Similarity-based navigation in visualized collections of historical documents. *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Lissabon, Portugal.

Besser, H. 2004. The past, present, and future of digital libraries. In Schreibman e.a. (reds.) 2004.

Bingham, A. 2010. The digitization of newspaper archives: Opportunities and challenges for historians. *Twentieth Century British History*, 21(2):225–31.

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

Boiangiu, C., A. Spataru, A. Dvornic, en I. Bucur. 2008. Merge techniques for large multiple-pass scanned images. *Proceedings of the 1st WSEAS International Conference on Vizualisation, Imaging and Simulation*.

Bor Ng, K. en J. Kucsma (reds.). 2010. Digitisation in the real world: Lessons learned from small and medium-sized digitization projects. New York: Metropolitan New York Library Council.

Borgman, C. L. 2010. The digital future is now: A call to action for the humanities. *Digital Humanities* Quarterly, 1–30.

Bunke, H. en A.L. Spitz (reds.). *DAS06.* Berlyn: Springer Verlag.

Cox, J. 2010. Academic libraries in challenging times. *An Leabharlann: The Irish Library*, 19(2):7–13.

Craig, H. 2004. Stylistic analysis and authorship studies. In Schreibman e.a. (reds.) 2004.

Crane, G., A. Babeu en D. Bamman. 2007. eScience and the humanities. *International Journal on Digital Libraries*, 7:117–22.

Deegan, M. en S. Tanner. 2004. Conversion of primary sources. In Schreibman e.a. (reds.) 2004.

Frischer, B. 2009. Art and science in the age of digital reproduction: From mimetic representation to interactive virtual reality. *I Congreso Internacional de Arqueología e Informática*, Sevilla.

Gatos, B., S. Mantzaris, S. Perantonis en A. Tsigris. 2000. Automatic page analysis for the creation of a digital library from newspaper archives. *International Journal on Digital Libraries*, 3:77–84.

Hajič, J. 2004. Linguistics meet exact sciences. In Schreibman e.a. (reds.) 2004.

Hockey, S. 2004. The history of humanities computing. In Schreibman e.a. (reds.) 2004.

Hodge, G. en Frangakis, E. 2004. *Digital preservation and permanent access to scientific information: The state of the practice*. Verslag geborg deur The International Council for Scientific and Technical Information (ICSTI) en CENDI.

Holley, R. 2007. Australian newspapers digitisation program: Helping communities access and explore their newspaper heritage. *Australian Media Traditions Conference*, Charles Sturt University, Bathurst.

Horton, T., K. Taylor, B. Yu en X. Xiang. 2006. "Quite right, dear and interesting": Seeking the sentimental in nineteenth century American fiction. *Digital Humanities*, 81–2.

Hota, S., S. Argamon, M. Koppel en I. Zigdon. 2006. Performing gender: Automatic stylistic analysis of Shakespeare's characters, *Digital Humanities*, 82–3.

Hu, C., A. Rose en B.B. Bederson. 2009. Locating text in scanned books. *Proceedings of the 2009 Joint International Conference on Digital Libraries*, Austin, Texas.

Hutcheon, L. 2006. *A theory of adaptation.* Londen: Routledge.

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

Jessop, M. 2008. Digital visualization as a scholarly activity. *Literary and Linguistic Computing*, 23(3):281–93.

Kae, A. en E. Learned-Miller. 2009. Learning on the fly: Font-free approaches to difficult OCR problems. *10th International Conference on Document Analysis and Recognition*, Barcelona.

Liebenberg, O. 2009. Persoonlike korrespondensie: e-pos, 27 Julie.

Lynch, C.A. 2008. The institutional challenges of cyberinfrastructure and e-research. *EDUCAUSE Review*, 43(6). http://www.uh.cu/static/documents/TD/The%20Institutional%20Challenges.pdf (9 Desember 2010 geraadpleeg).

Malhan, I. 2006. Trends of building and accessing digital collections and problems of digital divide in the emerging digital era. *Sri Lankan Journal of Librarianship and Information Management*, 2(1):5–10.

Skokowski, P. 2010. *Data tsunami – 5 Exabytes of data created every 2 days?* http://www.accellion.com/blog/2010/08/data-tsunami-5-exabytes-of-data-created-every-2-days (9 Desember 2010 geraadpleeg).

Platt, A. 2010. Developing an institutional repository at Southern New Hampshire University: Year One. In Bor en Kucsma (reds.) 2010.

Pramod, S.K., V. Ambati, L. Pratha en C. Jawahar. 2006. Digitizing a million books: Challenges for document analysis. In Bunke en Spitz (reds.) 2006.

Ramsay, S. 2004. Databases. In Schreibman e.a. (reds.) 2004.

Rommel, T. 2004. Literary studies. In Schreibman e.a. (reds.) 2004.

Rothenberg, J. 1999. *Avoiding technological quicksand: Finding a viable technical foundation for digital preservation*. *Report to the Council on Library and Information Resources*. Amsterdam: European Commission for Preservation and Access.

Rothenberg, J. en T. Bikson, 1999. *Carrying authentic, understandable and usable digital records through time*. *Report to the Dutch National Archives and Ministry of the Interior*, RAND-Europe.

Schreibman, S., R. Siemens en J. Unsworth (reds.). 2004a. *A companion to Digital Humanities*. Oxford: Blackwell.

—. 2004b. The Digital Humanities and Humanities Computing: An introduction. In Schreibman e.a. (reds.) 2004.

Schwarte, A., C. Haccius, S. Steenbuck en 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*, Lissabon, Portugal.

Serfontein, S. 2011. Persoonlike korrespondensie: e-pos, 25 Januarie.

Singh, S. 2003. Digital library: Definition to implementation, *Ranganathan Research Circle*, 1–18.

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za

Thomas, W.G. 2004. Computing and the historical imagination. In Schreibman e.a. (reds.) 2004.

Thomas, S., S. Cramond, M. Emery en P. Scott. 2005. *The digital library: Current perspectives and future directions*. Adelaide: University of Adelaide Library.

Tynan, D. 2010. *Prepare for data tsunami, warns google CEO*. http://www.pcworld.com/article/202817/prepare_for_data_tsunami_warns_google_ceo.html?tk=hp_new (9 Desember 2010 geraadpleeg).

Volk, M., T. Marek en R. Sennrich. 2010. Reducing OCR errors by combining two OCR systems. *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Lissabon, Portugal.

Warwick, C. 2004. Print scholarship and digital resources. In Schreibman e.a. (reds.) 2004.

Willett, P. 2004. Electronic texts: Audiences and purposes. In Schreibman e.a. (reds.). 2004.

Wooldridge, R. 2004. Lexicography. In Schreibman e.a. 2004.

http://www.nuance.com. 2010. http://www.nuance.com/for-business/by-solution/document-imaging-and-scanning/index.htm (9 Desember 2010 geraadpleeg).

Eindnota

[1] Die Erfenisstigting se kontrak met die Departement Sport, Kuns, Kultuur en Ontspanning is nie verleng vir ‘n tweede jaar nie, maar ten tye van die skrywe van hierdie artikel was daar geen aanduiding dat dit die digitaliseringsprojek sal skaad nie.

ISSN 1995-5928 | Tel: 021 886 5169 | E-pos: akademies@litnet.co.za