Textretrieval
als middel ter ontsluiting
van historische teksten

Textretrieval
als middel ter ontsluiting
van historische teksten

Werkgroep Textretrieval

Dit is een interne publicatie van het
Instituut voor Nederlandse Geschiedenis.

Instituut voor Nederlandse Geschiedenis / Den Haag 1996

Inhoudsopgave

1. Inleiding

De doelstelling van het ING is het inhoudelijk ontsluiten van historische bronnen. Deze activiteit kan uitmonden in het aanbieden van gestructureerde informatie, bijvoorbeeld in de vorm van inhoudelijk toegelichte inventarissen. Zij kan ook resulteren in het uitgeven van historische teksten. Deze uitgaven hebben niet zelden betrekking op zeer omvangrijke tekstbestanden, oplopend – in druk verspreid over verscheidene delen – tot duizenden pagina's. Men moet daarbij denken aan de editie van besluiten, notulen en rapporten van bepaalde bestuurlijke organen, of brieven van bepaalde functionarissen.

In toenemende mate dringt de elektronische uitgave zich op als alternatief voor de uitgave in boekvorm. Waar het de gestructureerde informatie betreft zijn bij het ING enkele publicaties in voorbereiding die alleen nog digitaal als gegevensbestanden ter beschikking van het onderzoek zullen worden gesteld. Ten aanzien van bibliografische informatie zijn reeds verschillende (Nederlandse) bibliografieën on-line raadpleegbaar; het *Repertorium* zal binnen afzienbare termijn dit voorbeeld volgen. Projecten waarbij informatie op een gestandaardiseerde wijze aan bronnen wordt onttrokken lenen zich eveneens zeer goed voor een digitale presentatie. Zo zal van het project *Rapporten van de Centrale Inlichtingendienst 1919-1940* het gegevensbestand op cd ROM worden gepubliceerd. Met behulp van deze database zal de gebruiker de rapporten op een aantal basisgegevens, inclusief een beknopte inhoudsaanduiding, kunnen doorzoeken en daarbij direct kunnen doorschakelen naar de afbeelding van de rapporten. Voor de editie van lopende tekst is het elektronische alternatief echter nog in statu nascendi. Toch heeft de noodzaak tot bezinning op dit alternatief zich de laatste jaren sterk opgedrongen. Een enkel woord ter toelichting is daarom op zijn plaats.

Zoals gesteld kent het ING projecten die de ontsluiting van zeer omvangrijke tekstbestanden tot onderwerp hebben. Voorbeelden zijn – ons beperkende tot de Nieuwe Geschiedenis – de editie van de notulen (‘acta’) van kerkelijke vergaderingen, van particuliere aantekeningen gemaakt tijdens de officiële bestuurlijke vergaderingen van de Staten van Holland, van de besluiten (‘resoluties’) van de Staten-Generaal en van de rapporten van de Gouverneur-Generaal in Nederlands-Indië aan de Heren XVII in Amsterdam. Aan de realisering van deze tekstedities is een fundamenteel probleem verbonden – dat bij de ene overigens zwaarder weegt dan bij de andere –, namelijk hoe de editie, gegeven de enorme omvang van het materiaal, tegen een redelijk te achten investering in tijd, personeel en financiële middelen te verwezenlijken is. Voorkomen moet immers worden dat vanwege schaarste aan middelen deze belangrijke ontsluiting van bronnenmateriaal stagneert of zelfs wordt gestaakt.

De bezinning op dit vraagstuk heeft nog niet tot conclusies geleid. Wel zijn uitgangspunten en mogelijke oplossingen besproken. Zo’n uitgangspunt is bijvoorbeeld dat als men eenmaal overgaat tot de ontsluiting van een archiefbestand, dat vanwege de centrale positie van het orgaan dat het bestand heeft voortgebracht van groot belang wordt geacht, die ontsluiting ook in principe alle informatie van de bron dient te omvatten. Organen als de Staten-Generaal, de VOC en de Staten van Holland namen op hun terrein een dergelijke centrale positie in en de weerslag daarvan is te vinden in archiefreeksen van besluiten, notulen of rapporten. Selectie op ‘relevantie’ of op een bepaald thema vormt ter ontsluiting van deze archiefbestanden geen realistische optie. Een mogelijk antwoord op de vraag hoe toch binnen redelijk te achten termen van inspanning een toereikende inhoudelijke ontsluiting zou kunnen worden verwezenlijkt, zou een (globale) index op het materiaal kunnen zijn, bij voorkeur met daaraan gekoppeld een afbeelding van de tekst zelf. Ook denkbaar is een digitale transcriptie van de tekst ter beschikking te stellen met het bieden van de mogelijkheid deze te doorzoeken op woorden en

op bepaalde categorieën van woorden. Kan op deze manier de functie van de conventionele index op zakelijke trefwoorden worden nagebootst? Deze vraag is het onderwerp van dit rapport.[1]

Over de ontsluiting van historische teksten met behulp van de computer zijn literatuur en voorbeelden voorhanden.[2] Die literatuur heeft vooral betrekking op inhoudsanalyse. Onderzoekers hebben de computer te hulp geroepen om de frequentie van bepaalde begrippen, hun inhoud en associatie met andere begrippen te helpen nagaan.[3] Minder aandacht is geschonken aan de ontsluiting van teksten en de specifieke problemen die zich daarbij kunnen voordoen. Daarbij wordt hier vooral gedacht aan het feit dat oudere historische teksten – van vóór ca. 1800 – gekenmerkt worden door een aanzienlijke variatie in gebruikte terminologie en spelling, en in het algemeen een tamelijk laag abstractieniveau waar het de gebezigde begrippen betreft. Het laatste probleem wordt in de conventionele ontsluiting van teksten gepoogd te ondervangen door 'handmatig' ontwikkelde indices waarin (gedeeltelijke) abstrahering van de in de oorspronkelijke tekst gebruikte terminologie plaatsvindt.

De werkgroep die het voorliggende rapport heeft opgesteld kreeg de taak recente tekstgeoriënteerde software te toetsen op bruikbaarheid bij het doorzoeken van oudere teksten. Uiteraard zijn de resultaten ook relevant voor de ontsluiting van teksten van recentere datum. De opdracht van de werkgroep en haar samenstelling is als bijlage bij dit rapport opgenomen. De werkgroep heeft gebruik gemaakt van het veel gebruikte tekstontsluitingssysteem ZyIndex, een product van M.S.C. Information Retrieval Technologies. Dit programma heeft de algemene kenmerken van textretrieval-software.[4] Dat wil zeggen dat het programma een uitgebreid scala aan zoektechnieken mogelijk maakt, zoals het zoeken:

- op woorden in de tekst;
- op varianten van woorden via 'wildcards' (door werk* in te voeren vindt men ook werken, werkloosheid, enz.);
- op spellingvarianten via 'fuzzy' zoeken;
- op combinaties van woorden via booleaans zoeken (AND OR NOT);
- op nabijheid van woorden, zodat gezocht kan worden naar het voorkomen van woorden in elkaars omgeving, waardoor contextgerichte zoekacties ondernomen kunnen worden;
- op grond van tevoren gedefinieerde synoniemen, woorden die dezelfde of een vergelijkbare betekenis hebben;
- op vaste structuren die in de documenten zijn aangebracht waardoor een zoekactie tot een bepaald deel van de tekst kan worden beperkt;
- naar concepten, waardoor met behulp van lange of complexe zoekopdrachten naar onderwerpen kan worden gezocht.

Het programma ZyIndex heeft op sommige onderdelen ook specifieke hoedanigheden, die in een enkel geval nadelen bleken te zijn. Zo bleek het met ZyIndex onmogelijk de woordenlijst te printen, hetgeen vanwege de essentiële functie van de woordenlijst onhandig is. Een groot manco is dat wanneer bij booleaans zoeken een woord met behulp van NOT wordt uitgesloten, het hele bestand waarin dat woord voorkomt wordt uitgesloten, ook als het wel gezochte woord

1 Bezinning op editiemethodieken in A.J. Veenendaal en J. Roelevink (eds.), *Unlocking government archives of the early modern period* (Den Haag 1995) en B.G.J. de Graaff en A.J. Veenendaal (eds.), *Bronontsluiting voor de negentiende en twintigste eeuw* (Den Haag 1995; interne publicatie ING).

2 Zie met name H. Voorbij, 'Analyse en ontsluiting van teksten met behulp van de computer', in O. Boonstra, L. Breure, P. Doorn (eds.), *Historische informatiekunde. Inleiding tot het gebruik van de computer bij historische studies* (Hilversum 1990) 130-183.

3 L. Breure, 'Tekstgerichte computertechnieken', in G. Rooijakkers e.a., *Trend of toekomst. Het gebruik van de computer in de geschiedwetenschap* (Nijmegen 1985) 65-72.

4 Zie bijv. J.C. Scholtes, 'Eigenlijk niets nieuws onder de zon bij fulltext retrieval techniek', in *VIP*, oktober 1994, 44-47.

in dat bestand voorkomt. Zo sloot de zoekopdracht 'geluk*NOT gelukken' alle bestanden uit waarin het woord 'gelukken' voorkomt, ook als het bestand het woord 'geluk' bevatte. Daardoor bleek de zoekfunctie NOT eigenlijk niet bruikbaar. Binnen het hierboven geschetste basisstramien van textretrieval-software verschillen de zoekprogramma's op specifieke kenmerken als deze. De keuze van een bepaald textretrieval-programma dient dus zeer weloverwogen te geschieden.[5] Ten behoeve van dit verkennende rapport werd een dergelijke voorstudie onnodig gevonden en zij is daarom achterwege gebleven.

Dit rapport kent de volgende bijlagen:

Bijlage I: Taakomschrijving en samenstelling van de werkgroep;
Bijlage II: Relevante literatuur
Bijlage III: De definiëring van concepten: het voorbeeld van het Dagverhaal van de Nationale Vergadering

[5] De Vereniging van Gebruikers van Online Informatiesystemen (VOGIN) zal begin 1997 een vergelijking van ca. 50 retrieval-softwarepakketten publiceren.

2. Drie tekstedities en textretrieval

2.1. De drie tekstedities en hun middelen tot ontsluiting

De werkgroep heeft gebruik gemaakt van de volgende drie tekstedities uit de *RGP*:
1. *Classicale acta van de Nederlandse Hervormde Kerk 1573 – 1620*;
2. *Particuliere notulen van de Staten van Holland 1620 – 1640*;
3. *Staatsregelingen 1796 – 1806*.

Deze tekstuitgaven hebben de volgende kenmerken:

ad 1. De editie van de classicale acta bestaat uit tekst waarin normalisering heeft plaatsgevonden van de leestekens en hoofdletters. Voor het overige is de 16e- en 17e-eeuwse spelling volledig gehandhaafd. Om het zoeken door de lezer te vergemakkelijken zijn de volgende elementen toegevoegd:
- een algemene *index op personen, aardrijkskundige namen en zaken*. Deze index dient tevens om verbasterde aardrijkskundige namen en persoonsnamen te identificeren en bij personen de volledige naam en zo mogelijk de functie toe te voegen;
- een *lijst van gemeenten* met de predikanten en een *lijst van predikanten* met de gemeenten die zij hebben gediend;
- in de tekst zijn de *vergaderingen gedateerd en per jaar genummerd*. Ook de *afzonderlijke artikelen hebben een nummer gekregen*;
- een *systeem van kopjes boven elk artikel* die via trefwoorden de lezer een (visueel) houvast bieden;
- zonodig een *concordantie*.

ad 2. In de editie van de particuliere notulen van Holland is de tekst aan de hand van een aantal vaste regels genormaliseerd. De editie is toegankelijk gemaakt via:
- een *algemene index op personen, aardrijkskundige namen en zaken*. Ook in deze editie dient de index tevens om verbasterde aardrijkskundige namen en persoonsnamen te identificeren en om bij personen de volledige naam en zo mogelijk de functie toe te voegen
- de tekst is geleed in *zittingen* en daarbinnen in de gedateerde *dag(delen)* waarop de Staten van Holland bijeen kwamen;
- de samenhangende *tekstonderdelen zijn genummerd*;
- elk tekstonderdeel kent een doorverwijzing naar het nummer van het vorige en het volgende samenhangende tekstonderdeel;
- met behulp van een *letter* is aangegeven van wiens hand de twee notulenteksten ('Stellingwerff en Schot') afkomstig zijn;
- een aanduiding met een *letter* (G) dat een tekstonderdeel ook in de gedrukte resoluties van de Staten van Holland voorkomt.

ad 3. De benutting van de editie Staatsregelingen is beperkt tot de daarin opgenomen letterlijke weergave van notulen van vertegenwoordigende en uitvoerende organen, met name het Dagverhaal van de Nationale Vergadering. Toegangen op de editie zijn:
- drie *indexen op persoonsnamen, zaken en geografische namen*;
- *concordanties* van artikelen van de verschillende staatsregelingen;
- *lijsten* van presidenten en leden van vergaderingen;
- *lijsten* van werkzaamheden van vergaderingen.

Ten aanzien van deze toegangen moet onderscheid worden gemaakt tussen de toegangen die reeds onderdeel van de oorspronkelijke tekst vormen en de toegangen die ter beter begrip van de tekst door de bewerker zelf zijn aangemaakt. Zo vindt in eigentijdse notulen al doorgaans structurering van de tekst plaats door deze te geleden op bijvoorbeeld de datum van een vergadering. Lijsten, zoals die van predikanten en gemeenten in de editie van de classicale acta, zijn door de bewerker van de bron samengesteld, deels zelfs op grond van materiaal dat niet in de bron zelf voorkomt.

Deze teksten, die digitaal op het ING voorhanden zijn, vormden het 'oefenmateriaal' voor het zoekprogramma, terwijl de bestaande toegangen in de gedrukte versie als referentiekader hebben gediend bij de beoordeling van de plus- en minpunten van de gebruikte textretrieval-software.

2.2. Omvang van de gebruikte bestanden

Zoals in hoofdstuk 1 aangegeven is de doelstelling van dit rapport de waarde van textretrieval-programma's ter ontsluiting van (ook) zeer omvangrijke tekstbestanden te meten. De gebruikte bestanden hadden de volgende omvang:

- classicale acta: 900 blz.;
- particuliere notulen: 600 blz.;
- staatsregelingen: 575 blz.

2.3. Zoeken

woord-frequentielijst

ZyIndex kent een woordfrequentielijst die direct inzicht geeft in de voorkomende woorden en spellingvarianten en daarmee dus in het taalgebruik. De onderzoeker kan zelf een woord opgeven om na te gaan of en op welke plaats in de tekst het woord voorkomt. Als zoekmiddel op afzonderlijke woorden is de lijst nuttig. Een handicap bij het systematisch zoeken vormt het grote aantal woorden waarop gezocht moet worden en de vele spellingvarianten in de twee vroegste teksten. Als (eerste) middel om na te gaan welke woorden in de tekst voorkomen is de woordenlijst onmisbaar.

ruiswoordenlijst

Het gebruikte zoekprogramma kent een op modern Nederlands geschoeide standaardruis-woordenlijst, bestaande uit woorden die niet in de woordfrequentielijst zijn opgenomen en daarom bij zoekopdrachten automatisch buiten beschouwing worden gelaten. Men moet hier bijvoorbeeld denken aan lid- en voegwoorden. Deze standaardlijst bleek voor de 16e- en 17e-eeuwse teksten niet goed bruikbaar, voor de vroeg 19-eeuwse beter, aangezien de Nederlandse taal toen dichter bij het hedendaagse Nederlands stond. Aanpassing van de ruiswoordenlijst door het toevoegen van typisch oudere woorden als 'ende' en 'sodat' is echter mogelijk. Bij toevoeging van woorden aan de ruiswoordenlijst moet overigens worden bedacht dat op deze woorden niet meer kan worden gezocht, zodat enige voorzichtigheid op haar plaats is.

fuzzy zoeken en wildcards

De 'fuzzyfunctie' van ZyIndex maakt het in principe mogelijk bij zoekopdrachten het probleem van spellingvarianten van bepaalde woorden te ondervangen. Er zijn oplopende 'fuzzy-graden', waarbij graad 1 een afwijking van één letter mogelijk maakt. Bij elk van de drie teksten bleek de ruis ook bij 'fuzzy-graad' 1 te groot te zijn (teveel 'missers') terwijl voor de hand liggende varianten eigenlijk een hogere 'fuzzy-graad' zouden vereisen. Zo werd met 'graad 1' van het

woord ‘onderwijs’ wel ‘onderwijl’ gevonden, maar niet ‘onderwys’ dat immers twee letters afwijkt.

Veel zinvoller bleek het werken met ‘wildcards’ te zijn, omdat hierbij kan worden aangegeven welke letters in een woord mogen afwijken. Kennis van taal en tekst en enige hersengymnastiek zijn voorwaarden om de ‘wildcards’ met vrucht aan te wenden. De toepassing van vooral de code * in een opgegeven woord leidt tot preciezere treffers dan de meer ongespecificeerde ‘fuzzy-functie’. Het gebruik van ‘wildcards’ leent zich bovendien het meest bij varianten van woorden met een unieke stam. Zo leek A*mst* alle varianten van Amsterdam/Aemstelredamme enz. te dekken. Veel spellingvarianten in 17-eeuwse teksten bleken aldus te kunnen worden ondervangen. Uiteraard leverde deze zoektechniek ook ongewenste zoekresultaten op, maar het aantal missers bleek niet hinderlijk hoog te zijn.

Men zou een vergelijking tussen deze twee zoektechnieken als volgt kunnen samenvatten. ‘Fuzzy’ zoeken vergt weinig denkwerk, maar levert veel missers op; het zoeken met ‘wildcards’ vereist veel denkwerk, maar kent een hoge mate van succes.

thesaurus (synoniemenlijst)

Het programma ZyIndex kent – wat het noemt – een moderne standaard-thesaurus. Deze zou beter een synoniemenlijst genoemd kunnen worden. De thesaurus bleek ter opsporing van synoniemen vooral tot hulp te zijn voor het zoeken in de notulen die in de edities van de staatsregelingen van na 1795 voorkomen. Voor toepassing op 16e- en 17-eeuwse teksten bleek de thesaurus niet bruikbaar. Met andere woorden, om een tekst te doorzoeken die dichterbij het hedendaagse Nederlands staat dan de 16e- en 17e-eeuwse teksten bleek de thesaurus goed te voldoen.

booleaans zoeken

Met de booleaanse operatoren AND, OR en NOT is het mogelijk naar combinaties van bepaalde namen of zakelijke trefwoorden te zoeken en daarbij bepaalde woorden uit te sluiten. Deze mogelijkheid werkt, indien de zoekopdracht niet al te uitgebreid is samengesteld, tamelijk eenvoudig. Voor oudere teksten is inpassing in de zoekopdracht van ‘wildcards’ absoluut noodzakelijk. Ook hier geldt dat een redelijke kennis van de tekst en de spellingvarianten een vereiste is, wil men tenminste gericht en met redelijk succes kunnen zoeken. Op het aan ZyIndex klevende probleem bij gebruik van de operator NOT is in paragraaf 1 al gewezen.

zoeken op nabijheid van woorden

Een belangrijk hulpmiddel is het zoeken op de nabijheid van bepaalde woorden. Die nabijheid – in aantallen woorden – kan door de onderzoeker zelf worden opgegeven. In bijvoorbeeld de verslagen van de vertegenwoordigende lichamen na 1795 werden met grote regelmaat requesten over een scala aan onderwerpen aan de orde gesteld. Met het zoeken op de nabijheid van woorden bleek het mogelijk het onderwerp waarop de requesten betrekking hadden – bijvoorbeeld belasting – af te zonderen van het totaal aantal ingediende requesten. Mutatis mutandis geldt deze zoekmogelijkheid ook voor de 16e- en 17e-eeuwse teksten. Deze zoekfunctie maakt het dus mogelijk zeer gericht naar een specifieke inhoudelijke context van een woord te zoeken.

concepten

Een textretrieval-programma is per definitie bedoeld om zoekopdrachten uit te voeren naar bepaalde in de tekst voorkomende woorden. Voor onderzoek naar woorden en woordverbanden, zoals wordt verricht door een taalkundig onderzoeker, is een textretrieval-programma zonder meer geschikt. De historisch onderzoeker is echter ook of nog meer geïnteresseerd in onderwerpen, in abstracties die als zodanig juist niet woordelijk in de tekst voorkomen. Zij zijn een constructie van hemzelf. In een conventionele index wordt dit probleem door de samensteller van de index ondervangen door begrippen als ‘financiën’ of ‘sociale zorg’ te definiëren en daarin concrete termen en hun synoniemen onder te brengen. Er wordt met andere woorden een hoger abstractieniveau over de eigenlijke tekst heengelegd.

Textretrieval-software kent de mogelijkheid om – wat ZyIndex noemt – 'concepten' te definiëren. Een concept is een zoekopdracht bestaande uit een reeks van in de tekst voorkomende woorden, waarbij tegelijkertijd gebruik kan worden gemaakt van andere zoekfuncties, zoals booleaans zoeken, zoeken op de nabijheid van bepaalde woorden, het afbakenen van de opdracht tot bepaalde eenheden van de tekst, zoeken met wildcards enz. Met deze methode kan gericht worden gezocht naar onderwerpen die breder zijn dan een enkel woord. Het is duidelijk dat naarmate het onderwerp ruimer is gedefinieerd en het concept dus ingewikkelder is, het moeilijker wordt de zoekopdracht (sluitend) te formuleren. Men zou ook kunnen stellen dat de onderzoeker dus afhankelijk is van de aard van de tekst. Hoe meer de eigentijdse auteur van de tekst – bijvoorbeeld een secretaris – gebruik maakt van vaste begrippen en bewoordingen, hoe eenvoudiger en succesvoller de gebruiker een onderwerpgerichte zoekactie zal kunnen ondernemen. Deze algemene constatering werd bevestigd bij het definiëren van concepten aan de hand van de gebruikte teksten.

De notulist van de classicale acta bezigde een veelheid aan concrete termen om zaken mee aan te duiden, terwijl van omvattender abstracte begrippen veel minder gebruik werd gemaakt. Een begrip als 'financiën' komt als zodanig niet in de tekst voor, hoewel tal van financiële zaken herhaaldelijk werden besproken. Om toch te zoeken naar het onderwerp 'financiën' zou een concept gedefinieerd moeten worden dat bestaat uit een lange opsomming van woorden, die aan financiële begrippen gerelateerd zijn, zoals: geld, penningen, (on)kosten, rekening, betaling, beurs, collecte, profijt, gage, tractement, boete, rentmeester en de in de tekst voorkomende munt-namen. Een dergelijke samengestelde zoekopdracht zou heel lang worden en tot de nodige ruis leiden. In dit geval zou het alternatief moeten zijn niet naar het onderwerp 'financiën' te zoeken, maar dit brede onderwerp op te delen in kleinere eenheden (zoals uitgaven, ontvangsten, munten).

Aan het eind van de 18e eeuw blijkt het begripsgebruik zich in een abstractere richting te hebben geëvolueerd. Dat is althans de conclusie die wordt getrokken uit de proef om aan de hand van concepten naar bredere onderwerpen te zoeken in het Dagverhaal van de Nationale Vergadering. 'Sociale zorg' gedurende de Bataafse Republiek bleek met een combinatie van de woorden armoede, armbestuur, armenfonds, armenhuis, aalmoezen en bedeling heel wel boven water te krijgen. Natuurlijk figureerden in de tekst van het Dagverhaal ook andere verwante woorden, zoals behoeftigen en gesticht, maar deze kwamen slechts zeer sporadisch voor en dan steeds in de nabijheid van een woord dat al wel in de zoekopdracht was opgenomen. Het aantal missers, te weten uitkomsten van de zoekactie die geen betrekking op het gezochte onderwerp bleken te hebben, was gering en deed zich met name voor bij zeer ruime zoekopdrachten die bestonden uit een groot aantal woorden uit de tekst. In bijlage III zijn enkele concepten uitgewerkt die gebruikt zijn voor het doorzoeken van het Dagverhaal.

Het zoeken naar concepten bleek derhalve met redelijk succes te kunnen worden uitgevoerd. Enkele problemen dienden zich aan. Ten eerste leidt het zoeken naar abstractere begrippen aan de hand van complexe zoekopdrachten ook tot niet bedoelde treffers. Gelijke woorden kunnen immers in verschillende betekenissen voorkomen. De ruis neemt dus toe. Ten tweede wordt het definiëren van zoekopdrachten moeilijker naarmate het niveau van abstractie van het gezochte onderwerp verder verwijderd raakt van het woordgebruik van de tekst. Hoe concreter en onsystematischer het woordgebruik, des te moeilijker, en hoe abstracter en consistenter des te eenvoudiger valt er te zoeken naar begrippen. Met de kroniekschrijver zal de onderzoeker in dit opzicht meer te stellen hebben dan met de filosoof.

Naast deze problemen zijn duidelijke voordelen aan te wijzen van het computergebruik. De onderzoeker is niet afhankelijk van de kwaliteit van de index op een tekst. Hij kan immers voortdurend, al naar gelang zijn vraagstelling, tot formulering van nieuwe vragen en nieuwe

zoekacties overgaan. De mogelijkheid verschillende woorden in één zoekopdracht te vangen, levert resultaten op die met een conventionele index alleen met verschillende na elkaar uitgevoerde zoekacties behaald zouden kunnen worden. Concepten die door de gebruiker zelf zijn samengesteld, zijn tot slot beter verifieerbaar dan een conventionele index waarvan achteraf niet goed valt na te gaan welke woorden in algemenere begrippen zijn samengevat.

conclusie

De zoekfuncties van het textretrieval-programma en de zoekstrategie die op basis van de zoekfuncties valt te ontwikkelen zijn een reëel alternatief voor de in conventionele (gedrukte) uitgaven voorkomende indices. Het zoekprogramma biedt aan de ene kant meer. De onderzoeker heeft met de automatische zoekfuncties een veel groter bereik met betrekking tot het aantal in de tekst voorkomende woorden dan hij met een conventionele index ooit kan hebben. Zij maken het ook mogelijk sneller (grote) bestanden te doorzoeken dan met een conventionele index ooit mogelijk kan zijn. Laatstgenoemde dwingt de onderzoeker immers ter beantwoording van vragen van samengestelde aard (combinaties van bepaalde begrippen) de index verscheidene malen achter elkaar te hulp te roepen. De zoekmogelijkheden van een geautomatiseerd systeem zijn flexibeler dan die van een conventionele index, omdat de onderzoeker beter in staat is zelf zijn vragen te formuleren zonder afhankelijk te zijn van de kwaliteit van een index. Het zoekprogramma biedt aan de andere kant ook minder: abstracte begrippen die als zodanig niet in de tekst voorkomen kunnen minder gericht worden gedefinieerd dan in een conventionale index, die de mogelijkheid biedt de gebruiker in de richting van aanzienlijke abstracties te sturen (met subjectiviteit als bij-effect). Spellingvarianten en in de tekst voorkomende synoniemen kunnen bij gebruik van een textretrieval-programma een hindernis vormen. In de volgende paragraaf zal worden nagegaan hoe met aanpassingen van de zoekfuncties en met bewerkingen van de tekst de gebruiker verder kan worden geholpen.

2.4. Aanpassingen en bewerkingen

ruiswoordenlijst

Het is mogelijk de ruiswoordenlijst aan te passen aan de eigenaardigheden van een specifieke tekst. Dat zou kunnen door via steekproeven die eigentijdse woorden uit de tekst op te sporen die een historisch onderzoeker normaal gesproken niet bij zijn vraag zal betrekken. Deze woorden kunnen daarna aan de ruiswoordenlijst worden toegevoegd waarna deze bij zoekopdrachten buiten beschouwing worden gelaten.

velden ter verkleining van het zoekgebied

Door middel van het aanmaken van velden in de tekst kan deze worden geleed, bijvoorbeeld chronologisch naar de zittingen van vergaderingen of naar inhoudelijk samenhangende tekstonderdelen (die ook in een chronologische volgorde kunnen staan). In paragraaf 2.1. is er op gewezen dat de historische tekst zelf vaak al een 'veldindeling' kent die in de te digitaliseren tekst overgenomen zou kunnen worden. Uiteraard kan de bewerker ook los van de structuur van de oorspronkelijke tekst een indeling aanbrengen. De voordelen mogen in beide gevallen duidelijk zijn: op deze wijze kan het bereik van een zoekopdracht worden beperkt en door de verkleining van de zoekopdracht wordt het probleem van de omvang van teksten beheersbaar gemaakt. Het zal wel nodig zijn een optimum te bepalen, namelijk welke graad van onderverdeling als zinnig of daarentegen storend wordt ervaren.

Het lijkt evenzeer zinnig om persoons- en aardrijkskundige namen afzonderlijk te markeren. Langs deze weg genereert men indices waarin kan worden gebladerd. Ook kunnen door deze markeringen van persoons- en geografische namen specifieke zoekacties worden ondernomen naar deze typen gegevens en kunnen zij natuurlijk ook van zoekopdrachten worden uitgesloten.

Via een noot of hypertekst zou eventueel de identificatie van personen en plaatsnamen kunnen worden getoond.

annotatie

Annotatie kan op verschillende manieren aan de digitale tekst worden toegevoegd. Zij kan als voet- of eindnoten worden aangebracht. Ook is het denkbaar (en functioneler) de noot door middel van hypertekst toe te voegen, zodat door middel van klikken op een woord of een nootnummer de noottekst in een apart klein kadertje verschijnt.

thesaurus

Een eigentijdse synoniemenlijst (in ZyIndex thesaurus genoemd) is zonder meer nuttig. Indien via deze lijst de gebruiker wordt gewezen op de verschillende termen die in de tekst voor eenzelfde begrip worden gebruikt, worden het zoeken op basis van woorden en het samenstellen van concepten sterk verbeterd. Het aanleggen van een dergelijke synoniemenlijst kost uiteraard de nodige tijd.

Ook een andere bewerking is zinvol. Zo kunnen spellingvarianten hinderlijk zijn als deze verschillende beginletters hebben. Voorbeelden zijn ‘so’ en ‘zo’ of ‘clock’ en ‘klok’. Door bijvoorbeeld (automatische) ‘zie-verwijzingen’ in de synoniemenlijst aan te brengen kan de onderzoeker op het rechte spoor worden gebracht. Steekproeven zouden misschien in dit geval afdoende kunnen zijn. Een stap verder is de synoniemenlijst zo in te delen dat een thesaurus ontstaat van woorden die gegroepeerd zijn in abstractere begrippen of categorieën. Op die manier zou ook geautomatiseerd naar die abstracties kunnen worden gezocht. Indien deze thesaurus beperkt zou blijven tot één bepaalde tekst is het verschil met het samenstellen van concepten niet wezenlijk. Een thesaurus werpt wellicht vooral vruchten af als deze bedoeld is ter raadpleging van een groot aantal verschillende, maar nauwe samenhang vertonende, teksten. De thesaurus zou dan dienst kunnen doen als algemeen (geautomatiseerd) verwijzingssysteem naar synoniemen van woorden en naar begrippen. Het opbouwen van een thesaurus met deze functie zal een veeleisend karwei zijn dat een aanzienlijke tijdsinvestering kost en overigens het beste ‘al werkende’ kan worden ontwikkeld.

conclusie

De vermelde aanpassingen en bewerkingen maken het mogelijk allerlei ontsluitingsmiddelen toe te voegen aan de standaard zoekfuncties. Zo men wil kan de digitale editie op eenzelfde wijze worden ontsloten als een gedrukte editie (zie par. 2.1 ten aanzien van de drie in druk verschenen tekstedities). Hoe ver de uitgever moet of wil gaan om de gebruiker met speciale hulpmiddelen terzijde te staan is afhankelijk van:

- de aard van een uit te geven tekst;
- de gebruiksvriendelijkheid van het programma die wordt nagestreefd;
- de investering die de uitgever wil doen in tijd, personeel en financiële middelen.

Op de keuzen die zich in dit verband aandienen wordt in de volgende paragraaf ingegaan.

3. Keuzen bij de editie van elektronische teksten

In de inleiding is uiteengezet dat de gedachte elektronische hulpmiddelen te gebruiken bij de ontsluiting van teksten met name is ingegeven door de hoop dat daarmee zeer omvangrijke tekstbestanden van duizenden tot tienduizenden bladzijden tekst toegankelijk kunnen worden gemaakt. De vraag luidde hoe bij een aanvaardbare inzet van tijd, personeel en financiële middelen op bestanden van dit volume een toereikende inhoudelijke ingang kan worden gemaakt. De conventionele tekstuitgave vergt teveel van de beschikbare middelen. Biedt een textretrieval-programma een werkbaar alternatief?

Een voorwaarde lijkt te zijn dat de tekst voldoende uniformiteit kent in taal, idioom en inhoud. Hoe diverser immers de tekst in deze drie opzichten zal zijn, hoe sterker zich de eis zal opdringen allerlei aanpassingen en bewerkingen toe te voegen om die uniformiteit zelf te creeëren. Van de drie gebruikte teksten bleken de notulen uit de Bataafs-Franse Tijd meer aan die voorwaarde te voldoen dan de 16e- en vroeg 17e-eeuwse acta van de kerkelijke classes. In de Bataafse bronnen kwamen minder spellingvarianten en synoniemen voor. Dit leidde tot minder problemen bij het formuleren van samengestelde zoekopdrachten. De gebruikte oude teksten vergen in dit opzicht een grotere bewerking om de gewenste mate van toegankelijkheid te kunnen scheppen. In het algemeen zullen ambtelijke stukken, zoals notulen, geschikter zijn voor ontsluiting via textretrieval dan verhalende of persoonlijke bronnen.

Ook de gebruiksvriendelijkheid stelt eisen aan de wijze van ontsluiting. Professionele onderzoekers zijn sinds meer dan een eeuw vertrouwd met het type ontsluitingsmiddelen als in 2.1 genoemd. De overgang naar de benutting van een geautomatiseerd systeem van tekstontsluiting zal nieuwe heuristische vaardigheden van de historicus vergen. Die overgang kan niet abrupt zijn. Daarom zal de bewerker doorgaans kunnen volstaan met een louter technische handleiding bij het zoekprogramma. Hij zal tenminste aan de hand van uitgewerkte voorbeelden wenken moeten geven hoe het programma gebruikt kan worden bij het doorzoeken van een bepaalde tekst. Men kan daarbij met name denken aan formuleren van complexe zoekopdrachten. De bezorger van de bron kan nog een stap verder wordt gaan door de gebruiker terwille te zijn door hulpbestanden als de ruiswoordenlijst en synoniemenlijst aan te passen, subeenheden aan te brengen en een aanzienlijke voorraad van concepten te ontwerpen. Deze handreikingen staan creativiteit van de gebruiker in dit opzicht overigens niet in de weg. Integendeel, automatische zoekprogramma's hebben het grote voordeel dat de onderzoeker zelf zoekopdrachten kan formuleren en herformuleren en deze met grote snelheid door het programma kan laten beantwoorden.

De noodzakelijke inzet van personeel en financiële middelen valt te splitsen in werkzaamheden die een conditio sine qua non zijn voor de elektronische uitgave en die activiteiten die men tot de categorie 'verrijking' zou kunnen rekenen.

Uiteraard dient de tekst digitaal beschikbaar te zijn. In het geval van handschriften zal een transcriptie nodig zijn. Deze kunnen door een niet-wetenschappelijke kracht worden gemaakt en moeten door een wetenschappelijk medewerker worden gecontroleerd. Een genormaliseerde transcriptie zou door de eliminatie van spellingvarianten een voordeel zijn boven een diplomatische transcriptie, maar uiteraard dienen de wetenschappelijke eisen de overhand te hebben boven zulke praktische overwegingen. Bij gedrukte teksten kan volstaan worden met scannen of handmatige invoer van de tekst.

Tot alle overige werkzaamheden kan desgewenst worden besloten. De indeling van de tekst in velden, bijvoorbeeld in zittingen van vergaderingen, en markeringen, van bijvoorbeeld persoons- en plaatsnamen, kunnen worden aangebracht door niet-wetenschappelijk personeel

aan de hand van instructies. De opbouw of aanpassing van een bestaande synoniemenlijst is een wetenschappelijke activiteit, evenals het identificeren van persoons- en plaatsnamen, het samenstellen van de annotatie en het formuleren van concepten. Men kan zich voorstellen dat bij teksten die aan de genoemde criteria van uniformiteit voldoen (een zekere eenheid in taal, idioom en inhoud) de bewerker zich ten aanzien van de formulering van concepten tevreden kan stellen met het uitwerken van een aantal voorbeelden. Immers, met hulp van die voorbeelden kan de onderzoeker zijn eigen concepten bepalen. Concepten hebben overigens potentieel een cumulatieve werking; eenmaal geformuleerd zijn deze op de tekst als geheel toepasbaar. Daarin ligt een niet te onderschatten tijdswinst voor de bewerker.

Op grond van deze drie invalshoeken – aard van de tekst; eisen die de gebruiker stelt; beschikbare middelen – dient een keuze of 'mix' te worden gemaakt uit de volgende (niet limitatief genoemde) middelen tot ontsluiting:

a. zoekprogramma zonder meer;
b. indeling van de tekst in subeenheden (zoals chronologische en/of inhoudelijk samenhangende tekstonderdelen);
c. markering in de tekst van bepaalde woorden (zoals persoons- en geografische namen);
d. aanpassing van een hulpbestand, zoals de ruiswoordenlijst, aan de eigenaardigheden van de tekst;
e. ontwerp van concepten, bij wijze van voorbeeld of bij wijze van systematische bewerking;
f. ontwerp van een synoniemenlijst, en eventueel een thesaurus van abstracte begrippen;
g. handleiding, waarvan de inhoud en uitgebreidheid afhankelijk zijn van de gekozen 'mix' van middelen tot ontsluiting.

Alvorens een voorstel te doen voor mogelijk te starten (proef)projecten, worden beknopt enkele voorbeelden besproken van bestaande 'full-text' cd ROMs op historisch gebied.

4. De zoekmogelijkheden in enkele historische 'full-text' cd's

4.1. Algemeen

Wanneer men de beschikbare titels van full-text cd's beziet, dan valt het op dat met name academische uitgevers als Oxford University Press en Chadwyck-Healey actief zijn op dit terrein. Zij publiceerden titels als *Jane Austen, Riverside Shakespeare* (Oxford University Press) en *Patrologia Latina, Voltaire electronique* (Chadwyck-Healey). In het algemeen zijn overzichten van de beschikbare titels veelal te vinden bij universiteitsbibliotheken (zie bijvoorbeeld het overzicht op Internet op het adres http://www.library.yale.edu/pubstation/test/fulltext.html).

De werkgroep heeft gezocht naar cd's met teksten, van bij voorkeur historische aard, welke doorzoekbaar zijn door middel van een index. Om een goed beeld te krijgen zijn zowel wetenschappelijke cd's als commerciële cd's in deze studie betrokken. Onder de keuze vallen cd's voor het Ms-Windows besturingssysteem alsmede een enkele voor Ms-DOS. cd's voor de Macintosh zijn niet in beschouwing genomen. Daarnaast heeft de beschikbaarheid van de cd's (in dit geval op de R.U. Leiden) ook een rol gespeeld.

Op basis van deze verkenning zijn de volgende cd's geselecteerd:[6]

1. *ADMYTE* (*Archivo Digital de Manuscritos y Textos Espanoles*), een cd uitgegeven door het Spaanse ministerie van Cultuur in het jaar van de Quinto Centenario van 1492, met Spaanse Postincunabelen en manuscripten;
2. *Library of Christian Latin Texts CLCLT-2*, uitgegeven door het Centre de Traitement Electronique des Documents, Université Louvain-la-Neuve (CETEDOC);
3. *Patrologia Latina database*, een verzameling van zes cd's met daarop de digitale editie van Jacques-Paul Migne's *Patrologia Latina* (1844-1855, 1862-1865), met Latijnse literatuur van Tertulianus (ca. 155 – na 220) tot het jaar 1216;
4. *The Hartlib Papers*, een cd van de University of Sheffield met transcripties van de overgebleven manuscripten van Samuel Hartlib (ca. 1600-1662) alsmede de digitale foto's van dit materiaal;
5. *Chaucer, Life and Times* van Primary Source Media, bevattende *The Riverside Chaucer* in het Middel-Engels, alsmede een moderne vertaling.

4.2. Een vergelijking

voordelen van Windows boven DOS

Het nadeel van Ms-DOS programma's is dat er vrijwel nauwelijks sprake is van een standaard interface, met als gevolg dat je bij veel DOS programma's eerst de eigenaardigheden van de interface moet leren gebruiken. Dit geldt ook voor *Christian Latin Texts*, waarbij om maar eens een voorbeeld te noemen, de PgDn-toets moet worden gebruikt om van het ene veld naar het volgende te springen.

één cd of een aantal cd's

Op een cd ROM kan rond de 640 Mb aan gegevens worden geschreven. Voor alleen tekst is dat meestal ruim voldoende. Wil men echter ook de grafische afbeeldingen van de originele pagina's

[6] *ADMYTE Archivo digital de manuscritos y textos Espanoles* (Madrid: Micronet S.A., Madrid: Biblioteca Nacional/ Ministerio de Cultura 1992); *CETEDOC Library of Christian Latin Texts CLCLT-2* (Brepols 1994); *Patrologia Latina database* (Alexandria Va.: Chadwick-Healey 1995); *The Hartlib Papers: A Complete Text and Image Database of the Papers of Samuel Hartlib (c. 1600-1662)* (UMI, Ann Arbor Mich. 1995); *Chaucer, Life and Times* (Primary Source Media Ltd, Reading 1995).

(de 'images') met een redelijke kwaliteit opnemen, dan is deze 640 Mb al snel ontoereikend. Het alternatief is de keuze voor een aantal cd's. Dat heeft echter als nadeel dat de meeste pc's slechts beschikken over één cd-ROM drive, waardoor de gebruiker voortdurend cd's moet wisselen.

wetenschappelijke versus commerciële cd's

Bij wetenschappelijke uitgaven ligt de nadruk meestal op een goede verzorging van de tekst en uitgebreide zoekmogelijkheden. Toch kan het geen kwaad deze te vergelijken met commerciële cd's. Bij commerciële cd's zoals bijvoorbeeld *Chaucer, Life and Times* is veel meer zorg besteed aan het maken van een aantrekkelijke interface voor de gebruiker.

booleaans zoeken, wildcards en haakjes

De meeste indexen staan het gebruik van Booleaanse operatoren toe, veelal in combinatie met haakjes. Bij de *Chaucer*-cd is dit wel beperkt tot een AND en OR-combinatie van twee woorden. Anders staat het met de wildcards of jokers. Het gebruik van de '*' en '?' wildcards aan het begin van een woord is alleen mogelijk bij de *ADMYTE* en *Christian Latin Texts*. De *Patrologia Latina database* kent daarentegen het groeperingssymbool '[ ]', waarmee bijvoorbeeld de zoekaktie 'epist[ou]la' zowel op 'epistola' als op 'epistula' past.

zoeken op nabijheid van woorden

Zoeken op de nabijheid van woorden is mogelijk bij de *ADMYTE*, de *Christian Latin Texts*, de *Patrologia Latina database* en *Hartlib Papers*. De notatie verschilt per programma. Kiezen voor nabijheid met inbegrip van de in de zoekopdracht gegeven volgorde kan slechts bij *ADMYTE* en bij *Patrologia Latina*. Een voorbeeld van een zoektocht in *Christian Latin Texts* (carne* /3 christ*) + (haeret*, arian*), leidde tot 28 treffers. Bij de *Hartlib Papers* kan de scope van een zoekvraag worden beperkt door gebruik te maken van de strukturele eenheden in de tekst. Zo kan bij het zoeken op verschillende termen de beperking worden opgegeven dat deze woorden in dezelfde alinea of in dezelfde zin dienen te staan.

divers gebruik van zoeken op nabijheid van woorden

	ADMYTE
.2	gescheiden door max. 2 woorden
	Christian Latin Texts
/2	gescheiden door max. 2 woorden
%2	gescheiden door max. 2 woorden in de gegeven volgorde
	Patrologia Latina
within 2 words of	gescheiden door max. 2 woorden
within 2 words after	gescheiden door max. 2 woorden in de gegeven volgorde
within 2 words before	gescheiden door max. 2 woorden in de omgekeerde volgorde
	Hartlib Papers
same paragraph	zelfde alinea
same sentence	zelfde zin
same phrase	zelfde zinsdeel

concepten

Het gebruik van concepten valt eigenlijk alleen te constateren bij *ADMYTE* en bij de *Hartlib Papers*. *ADMYTE* biedt een glossary of lemmata, waarin woordvarianten van dezelfde stam opgezocht kunnen worden. Dat is bij het Middel-Spaans ook wel nodig gezien de grote variatie in spelling. Een woord als 'Medecina' kan ook worden gespeld als 'Melezina' en kan zelfs worden voorafgegaan door een Arabisch lidwoord: 'Almelezinar'. Een druk op de knop 'search

lemmata' plaatst de betreffende groepering in de zoekdialoog-box en daarna begint de zoekaktie. De notatie voor die groeperingen in de zoekopdracht bestaat uit de woordvorm voorafgegaan door een '='teken (bv. '=medecina'). De concepten blijken helaas niet te combineren: '=medecina.y =anima" levert geen treffers op, '=medecina.y anim*' wel. Hetzelfde geldt voor '=mecedina.y =sciencia' vs. 'mecedina.y scienci*'.

Bij de *Hartlib Papers* kan de gebruiker concepten en zoekopdrachten bewaren in de vorm van zogenaamde 'topics' (naar het indexeringsprogramma Verity Topic). Op de cd staat ook een aantal voorgedefinieerde 'topics' om als voorbeeld te dienen. Veelvuldig in het werk van Samuel Hartlib optredende topoi zoals bijvoorbeeld de chiliastische redenering dat de bekering van alle Indianen het 'Einde der Tijden' naderbij zou brengen, zijn als 'topic' gedefinieerd op de cd.

fuzzy zoeken

De mogelijkheid tot 'fuzzy' zoeken of het zoeken met een opgegeven onnauwkeurigheid is bij geen van de cd's aangetroffen. Het NHDA maakt momenteel ook een proefapplicatie met de correspondentie van Michael Bakunin, een uitgave van het IISG. Deze worden geïndexeerd met ZyIndex, waardoor ook kan worden gezocht met 'fuzzy' zoeken. Zo kunnen de verschillende versies van persoonsnamen (Franse, Duitse en getranslitereerde Russische) met één zoekaktie worden gevonden.

overzicht van de kenmerken

	ADMYTE	**Chr. Lat. Texts.**	**Patr.Lat.**	**Hartlib**	**Chaucer**
Jaar	1992	1994	1995	1995	1995
OS	Win	DOS	Win	DOS	Win
Index	TACT	onbekend	DynaText	Verity Topic	onbekend
Interface	(Win)	tekstuele menu's	DynaText	tekstuele menu's	Visual Basic
Booleaans	.and/.or/.not. y/.o-	'+' (and) ',' (or) '#'	and/or/not	and/or/not	and/or
Haakjes	ja	ja	ja	ja	nee
Wildcards	*/?	*/?	*/?/[]	*/?	geen
Proximity search	ja	ja	ja	ja	nee
Concepten	glossary of lem-	nee	nee	Topic's	nee
Fuzzy	nee	nee	nee	?	nee
Extra's	images	nee	noten	images	geluid

5. Aanbeveling voor een vervolgproject

De werkgroep besluit haar eigen zoektocht positief gestemd. Zij is ervan overtuigd geraakt dat met behulp van textretrieval-software een zinvolle ontsluiting van historische teksten valt te bereiken. Problemen zijn er, en deze hebben in het voorafgaande de revue gepasseerd. Zij zijn volgens de werkgroep overkomelijk. Daarom beëindigt de werkgroep haar rapport met de aanbeveling een vervolgproject te starten.

De werkgroep beveelt aan dat één of meer teksten worden uitgekozen ter ontsluiting met behulp van een tekst-georiënteerd zoekprogramma. Die teksten moeten deel uitmaken van de categorie bronnen die in de inleiding expliciet is genoemd als 'doelwit' van dit rapport: zeer omvangrijke tekstbestanden afkomstig van belangrijk geachte organen of functionarissen die juist vanwege de omvang van de teksten problematisch zijn geworden om op een conventionele wijze te ontsluiten. Daarbij kan worden gedacht aan de 'Resolutiën van de Staten-Generaal' uit de 17e eeuw en het 'Dagverhaal van de Nationale Vergadering' eind 18e eeuw. Beide bronnen hebben een centraal karakter voor hun tijd en zijn niet of slechts partieel ontsloten. Beide staan al lange tijd hoog op de prioriteitenlijst om ontsloten te worden: de Resolutiën van de Staten-Generaal maken deel uit van het huidige programma van het ING, op het Dagverhaal van de Nationale Vergadering is al enkele malen (buiten het ING) vergeefs gepoogd een ingang te maken.

Een ontsluitingsproject zou in fasen opgezet moeten worden. Dat houdt in dat eerst op basis van een proefbestand van bijvoorbeeld één jaar Resolutiën en/of één jaar Dagverhaal bepaald wordt welke 'mix' van elektronische ontsluiting wordt nagestreefd (zie hiervoor hoofdstuk 3). Van direct praktisch belang is de vraag of de tekst wordt voorbewerkt door deze in subeenheden in te delen of door bepaalde woorden in de tekst te markeren. Tegelijkertijd zou de keuze voor het te gebruiken textretrieval-programma moeten worden gemaakt. Weet men eenmaal welke voorbewerking van de tekst wenselijk wordt geacht, dan zou de tekst gedigitaliseerd moeten worden door transcriptie (bij handschriften) of door handmatige invoer of scannen (bij gedrukte teksten). Nadat de digitale tekst beschikbaar is kunnen nadere bewerkingen plaatsvinden, zoals het aanleggen van een synoniemenlijst en een voorraad van concepten. Tot slot dient een handleiding samen te worden gesteld.

De eerste fase zou een voorlopige publicatie als uitkomst moeten hebben die aan gebruikers zou moeten worden voorgelegd. Allicht zal deze consultatie tot een aantal bijstellingen leiden. Op basis van die bijstellingen zou voorbereiding van fase twee plaats moeten vinden, die zou bestaan uit dezelfde handelingen als in de eerste fase, maar nu na toevoeging van nieuwe tekst. De tweede fase wordt in deze opzet besloten met de publicatie van de gehele tekst, inclusief de apart te publiceren handleiding.

6. Epiloog

De werkgroep heeft haar rapport optimistisch besloten met een aanbeveling voor een vervolgproject, niet langer als 'pilot' maar als entree tot een publicatie. Dat wil uiteraard niet zeggen dat de weg naar een 'full-text' cd ROM van het ING is geplaveid. Er dienen zich verschillende vragen aan. Uiteraard allereerst vragen die samenhangen met de aard van de te ontsluiten teksten en de ontsluitingsmiddelen die toegepast zullen worden. Daarover handelt dit rapport. Ten tweede moet worden overwogen hoe de toepassing van tekstgeoriënteerde software past binnen het informatiseringsbeleid van het ING als geheel: bijvoorbeeld op welke doelgroep wordt gemikt en in dat verband welke voorkennis van tijd en tekst mag worden verondersteld. Ten derde is een oriëntatie op het beleid van 'zusterorganisaties' (ook in het buitenland) van belang; niet alleen vanuit een oogpunt van algemeen beleid, maar ook als referentiekader bij de ontsluiting als zodanig, bijvoorbeeld ter beantwoording van de vraag in hoeverre naar (internationale) standaardisatie moet worden gestreefd. Ten vierde vragen de vaardigheden van de toekomstige gebruikers om met deze nieuwe techniek om te gaan de aandacht; wat mag men als instituut eisen van zijn gebruikers en hoe kan worden bevorderd dat de gebruiker over de noodzakelijke heuristische vaardigheden beschikt? En tot slot, in het verlengde hiervan: welke aanpak is binnen het ING gewenst om zijn medewerkers de nodige theoretische kennis en praktische vaardigheden op te laten doen om de weg naar de digitale ontsluiting van historische teksten met succes af te leggen?

Werkgroep Textretrieval

Bijlage I
Pilotproject Text-Retrieval

Pre-ambule

Het ING is gespecialiseerd in het ontsluiten van historische teksten en het verrichten van het daarvoor vereiste (archief) onderzoek. Dit heeft niet zelden betrekking op zeer omvangrijke tekstbestanden (oplopend tot duizenden pagina's tekst). Hun inhoud is vaak divers (bijvoorbeeld: politiek-, economisch-, en cultuurhistorische informatie door elkaar). Oudere historische teksten – van vóór ca. 1800 – worden gekenmerkt door een aanzienlijke variatie in gebruikte terminologie en veel spellingvarianten. Historische teksten worden in boekvorm gepubliceerd, waarbij de tekst via indices, lijsten e.d. voor de gebruiker toegankelijk wordt gemaakt. Bij het proces van 'handmatige' indicering vindt (gedeeltelijk) abstrahering van de in de oorspronkelijke tekst gebezigde terminologie plaats om de gebruiker een hanteerbaar apparaat bij zijn onderzoek in handen te geven. Indicering op begrippen vergt een aanzienlijke investering in tijd op wetenschappelijk niveau. Recente tekstgeoriënteerde software kan een aanvulling op of een alternatief voor de conventionele ontsluiting en uitgave in drukvorm zijn. Om inzicht te verwerven in de mogelijkheden en problemen bij gebruik van deze tekstgeoriënteerde software wordt een werkgroep ingesteld waartoe het volgende wordt bepaald:

Doelstelling

- inzicht te verwerven in de bruikbaarheid van tekstgeoriënteerde programmatuur bij de ontsluiting van historische teksten, en als gereedschap bij het bronnenonderzoek.
- uitspraken te doen, zo mogelijk aan de hand van uitgewerkte voorbeelden, over de bruikbaarheid van de volgende ontsluitingstechnieken:
  1. het indexeren van een vrije tekst (al dan niet vergezeld van een 'handleiding' bij toepassing op een specifieke tekst);
  2. het aanleggen van een thesaurus en synoniemenlijst bij een vrije geïndexeerde tekst;
  3. het structureren van teksten met behulp van markeringen, waarbij belangrijke inhoudelijke elementen worden gethesaureerd.
- vergelijking van de nieuwe technieken met conventionele technieken (i.c. handgemaakte indices) bij uitgave in drukvorm;
- globaal aan te duiden welke werkzaamheden noodzakelijk zijn om tot de verschillende bewerkingsmogelijkheden te komen, welk niveau van personeel daarvoor ingezet moet worden en welk tijdsbeslag ermee gemoeid is;

Materiaal

- oudere historische teksten met als specifieke problemen: veel spellingvarianten, geringe mate van uniformering in gebruik van begrippen;
- de keuze wordt bepaald tot: teksten Classicale Acta, teksten Notulen Holland ('Schot/ Stellingwerff'), teksten Staatsregeling Bataafs-Franse Tijd (met name 'Dagverhaal Nationale Vergadering'). Van al deze teksten zijn computerbestanden op het ING aanwezig;
- zo mogelijk een beknopte verkenning van (literatuur over) elektronische uitgaven van historische teksten die ontsloten zijn met behulp van tekstgeoriënteerde programmatuur;

Werkwijze

– instelling werkgroep bestaande uit:
  ING:
  * drie medewerkers (I. Huysman, J. Roelevink, A. Verheusen) die alle inhoudelijke kennis bezitten t.a.v. één der genoemde teksten en voldoende ingevoerd zijn met het gebruik van computers om de genoemde technieken te kunnen toepassen. De medewerkers van het ING besteden in de periode september/oktober 1996 elk maximaal 10 werkdagen aan het pilotproject;
  NHDA:
  * NHDA-medewerker (H. van Mourik) met kennis van de te gebruiken software en de betreffende ontsluitingstechnieken.
– de werkgroep start haar werkzaamheden zo spoedig mogelijk en rapporteert aan de directeur van het ING uiterlijk 1 november 1996;
– als voorzitter van de werkgroep treedt D. Haks op;

Tekstontsluitingssysteem

– de werkgroep heeft op het ING en op het NHDA de beschikking over een tekstontsluitingssysteem, waarvoor het NHDA een voorstel doet;

Vervolg

– na ontvangst van de rapportage bereidt de directeur verder besluitvorming voor. Deze kan bestaan uit projectvoorstellen tot ontsluiting via textretrieval van gedigitaliseerde tekstbestanden (hetzij via transcriptie, hetzij via scanning/OCR).

D. Haks; 15-7-1996 (2e versie)

Bijlage II
Relevante literatuur over tekstanalyse en textretrieval van historische teksten

Bourlet, C., 'Content Analysis of French Medieval Charters', in: A. Gilmour-Bryson, *Computer Applications to Medieval Studies. Studies in Medieval Culture* 12 (Kalamazoo 1984) 63-80.

Bozzi, A. en Capelli, G., 'Automatic Lemmatization of Latin Texts', in: H. Best, E. Mochmann, M. Thaller, *Computers in the Humanities and the Social Sciences. Achievements of the 1980s Prospects for the 1990s. Proceedings of the Cologne Computer Conference 1988* (München, London, New York, Paris 1991) 373-378.

Bozzi, A. en Cappelli, G., 'The Latin Lexical Database and Problems of Standardization in the analysis of Latin Texts', in: F. Hausman, R. Hertel, I. H. Kropac, P. Becker, *Datennetze für die Historischen Wissenschaften? Probleme und Möglichkeiten bei Standardisierung und Transfer Maschinenlesbarer Daten* (Graz 1987) 28-45.

Breure, L., 'Tekstgerichte computertechnieken voor historici', in: G. Rooijakkers, R. Schönberger, M. Smits, R. v.d. Weyer, T.S.M. v. d. Zee, (editors), *Trend of toekomst. Het gebruik van de computer in de geschiedwetenschap*, (Nijmegen 1985). p. 45-65.

Breyer, G., Finzsch, N., Schaefer, J., Straeter, J., Wengler, F. et al., 'Computer-Aided Content Analysis and Soft Data in Historical Social Research. An Attempt to Find a pragmatic Solution', in: *Historical Social Research / Historische Sozialforschung* 15 (1990) 206-213.

Bruce, D., 'Towards the implementation of text and discourse theory in computer assisted textual analysis', in: *Computers and the Humanities* 27 (1993) 357-365.

Burnard, L., 'Primary to Secondary: Using the Computer as a Tool for Textual Analysis in Historical Research', in: P. Denley, D. Hopkin, *History and Computing* (Manchester 987) 228-233.

Busa, R., 'The annals of humanities computing: the index Thomisticus', in: *Computer and the humanities* 14 (1980) 83-90.

Coates-Stephens, S., 'The analysis and acquisition of proper names for the understanding of free text', in: *Computers and the Humanities* 26 (1992) 441-457.

Dautray, P. en P. Muller, 'Lemmatisation et analyse des textes', in: F. Bocchi, P. Denley, *Storia & Multimedia. Atti del Settimo Congresso Internazionale / Proceedings of the seventh International Congress – Association for History and Computing* (Bologna 1994) 261-268.

Dendien, J., 'Access to Information in a Textual Database: Access Functions and Optimal Indexes', in: I. Lancashire, *Research in Humanities Computing 1. Selected Papers from the ALLC/ACH Conference, Toronto June 1989* (Oxford 1991) 308-324.

Olsen, M., 'The language of enlightened politics: the Société de 1789 in the French Revolution', in: *Computers and the Humanities* 23 (1989) 357-364.

Olsen, M. en L.G. Harvey, 'Computers in intellectual history: lexical statistics and the analysis of political discourse', in: *Journal of interdisciplinary history* 18:3 (1988) 449-464.

Stenvert, R., *Syllabus Thesaurusbouw* (Vakgroep Computer & Letteren) (Utrecht 1989).

Voorbij, Hans, 'Analyse en ontsluiting van teksten met behulp van de computer', in: Onno Boonstra, Leen Breure, Peter Doorn ed., *Historische Informatiekunde. Inleiding tot het gebruik van de computer bij historische studies* (Hilversum 1992) 130-183.

Bijlage III
De definiëring van concepten: het Dagverhaal van de Nationale Vergadering

Er is een aantal concepten gedefinieerd, waarbij geprobeerd is zowel enkele abstracte als enkele meer concrete begrippen op te nemen. De bestaande index op de reeks 'Staatsregelingen', bewerkt door L. de Gou, is hierbij gebruikt als referentiekader. De werkwijze voor de abstracte en concrete begrippen was gelijk. Allereerst moesten de relevante woorden gevonden worden. Dit heb ik gedaan met behulp van de woordenlijst en door te kijken naar de context van al bekende woorden. Daarna zocht ik op ieder woord afzonderlijk (ook met 'fuzzy' zoeken), waarbij ik lette op spellingvarianten en andere, onjuiste betekenissen van de woorden. Vervolgens bepaalde ik bij elk woord waar de 'wildcards' moesten staan en keek ik naar mogelijkheden om onjuiste betekenissen van een woord uit te sluiten. Hierna voegde ik de woorden samen tot een zoekopdracht. Wanneer sprake was van een samengesteld begrip, dat uit meer woordgroepen bestond, zoals 'democratische republiek' of 'provinciale schulden', ging ik tot slot na wat de afstand tussen deze groepen moest zijn.

Er moest ook telkens gekeken worden naar de Franse woorden met betrekking tot de concepten. Dit leverde geen problemen op. Soms omvatten de Nederlandse woorden door het gebruik van 'wildcards' automatisch de Franse woorden. Zo geeft 'revolut*' zowel 'revolutie' als 'revolution'.

Al met al was de werkwijze een kwestie van zoeken naar de juiste woorden, kijken naar de teksten en uitproberen van verschillende zoekopdrachten. Het lijkt me dat de vaardigheden voor het maken van concepten vrij snel aangeleerd kunnen worden.
De zoekopdrachten leidden tot de volgende resultaten:

Concept:	*Treffers:*	*Waarvan missers:*
1. provinciale schulden	136	15
2. samensmelting gewestelijke schulden	52	0
3. agent(schappen)	91	3
4. agent van Binnenlandse Zaken	4	0
5. agent tot zuivering grondvergadering	15	0
6. democratische republiek	10	0
7. geluk	231	20
8. buitenlandse invloed	25	2
9. invloed van het volk	24	4
10. dagbladen	23	0
11. sociale zorg	16	1
12. revolutie	71	5
13. revolutie van 22 januari 1798	86	0
14. onderwijs en opvoeding	41	0

ad. 1 De zoekopdracht van het concept luidde: *(gewest* or departement* or provinc*) w/30 (financ* or schuld*)*. Het relatief grote aantal missers is te wijten aan de te ruime definitie van de zoekopdracht. De gebruikte woorden komen te vaak voor in elkaars buurt. Ook de afstand die tussen de twee verschillende woordgroepen mocht bestaan, was te groot.

ad. 2 De zoekopdracht luidde: *amalgame* w/10 schuld**. Uitbreiding van deze opdracht met 'samensmelting' en 'samenvoeging' leverde geen extra treffers op.

ad. 3 De zoekopdracht luidde: *agent**. De drie missers hadden voorkomen kunnen worden door niet in de bestanden van 1805 en 1806 te zoeken. (De agentschappen werden in 1801 opgeheven).

ad. 4 De zoekopdracht luidde: *agent* w/15 binnenland**.

ad. 5 De zoekopdracht luidde: *agent w/10 (grondvergadering* and zuivering*)*. De index van De Gou verwees maar naar één van deze treffers.

ad. 6 De zoekopdracht luidde: *democratisch* w/30 republi**.

ad. 7 De zoekopdracht luidde: *geluk* or welva* or welz* or not gelukken*. Dit werkte dus niet. Er was een aantal missers met het woord 'gelukken'. Deze probeerde ik op deze manier, met 'or not', uit te sluiten. ZyIndex sluit dan echter het hele bestand uit, waar het woord 'gelukken' in voorkomt.

ad. 8 De zoekopdracht luidde: *invloed w/15 (engel* or fran* or vreem* or buitenland*)*.

ad. 9 De zoekopdracht luidde: *invloed w/15 volk*. De missers hadden betrekking op het voorkomen van de twee woorden in elkaars nabijheid, zonder dat de bedoelde betekenis van toepassing was.

ad. 10 De zoekopdracht luidde: *haag* courant or couranten* or nieuwspapier* or nieuwsblad* or dagblad* or journal or weerlicht*. Deze woorden komen bijna niet in een ander betekenis voor, waardoor volstaan kan worden met alleen een opsomming van alle woorden die met het concept te maken hebben.

ad. 11 De zoekopdracht luidde: *armen* or armbestuur or armefondsen or armhuizen or armoede or bedeeld* or aalmoezen*. Het was niet nodig sommige andere woorden die hiermee te maken hadden, zoals 'gesticht' en 'behoeftigen', op te nemen, omdat deze maar één keer voorkwamen en dan stonden ze in de context van een ander woord, dat al wel in de zoekopdracht was opgenomen.

ad. 12 De zoekopdracht luidde: *revolut* or staatsgreep or staatsomwen* or omwen**.

ad. 13 De zoekopdracht luidde: *22 janu* or 22 janvier*.

ad. 14 De zoekopdracht luidde: *onderw*s or school* or wetenschap* or universiteit or hoogleeraar* or academ* or leraaren or opvoeding or curator**.

Het was vrij gemakkelijk het aantal missers te bepalen door gewoon alle treffers na te lopen. Moeilijker was het vast te stellen of er relevante passages misten. Ik heb dit gedaan door de treffers te vergelijken met de index van De Gou, waarbij ik 1805-1806 buiten beschouwing heb gelaten. In de meeste gevallen leverden de zoekopdrachten meer treffers op dan de index van De Gou.

In het algemeen leidden de abstracte begrippen tot meer missers dan de concrete. Dit wordt waarschijnlijk veroorzaakt door het gebruik van meer woorden, waardoor er ook meer kans bestaat op andere betekenissen van die woorden.

Gebruikers zouden in een eigen zoekopdracht bestaande concepten uit kunnen breiden. Ze moeten dan wel goed gewezen worden op de evaluatievolgorde van de booleaanse operatoren.

A. Verheusen