

Data Archiving and Networked Services



Eindrapport

Een verkenning van de gewenste functionaliteit van de krantendatabank ten behoeve van wetenschappelijk onderzoek

Den Haag, 31 maart 2008

Maarten Hoogerwerf, Laurents Sesink en Caroline Voorbrood
Data Archiving & Networked Services (DANS)
T 070 3494450
Anna van Saksenlaan 51
2593 HW Den Haag
versie: 1.0

Inhoudsopgave

| | |
|---|-----------|
| INHOUDSOPGAVE | 2 |
| SAMENVATTING | 3 |
| 1 OPDRACHT | 5 |
| 1.1 AANLEIDING | 5 |
| 1.2 DOEL EN OPZET VAN HET ONDERZOEK..... | 5 |
| 1.3 AFBAKENING EN BEREIK | 6 |
| 1.4 BEOOGDE RESULTATEN | 6 |
| 2 VERLOOP VAN DE VERKENNING | 7 |
| 2.1 SELECTIE VAN ONDERZOEKERS | 7 |
| 2.2 TECHNISCHE MOGELIJKHEDEN KORTE TERMIJN..... | 7 |
| 2.3 TECHNISCHE MOGELIJKHEDEN LANGE TERMIJN..... | 7 |
| 2.4 ONDERZOEKSVRAGEN EN GEWENSTE FUNCTIONALITEIT..... | 8 |
| 2.5 WORKSHOP..... | 8 |
| 2.6 TERUGKOPPELING..... | 8 |
| 3 ONDERZOEKERS EN HUN ONDERZOEKSVRAGEN | 9 |
| 4 GEBRUIKERSFUNCTIONALITEIT | 12 |
| 4.1 KORTE TERMIJN..... | 12 |
| 4.2 MIDDELLANGE TERMIJN..... | 12 |
| 5 FUNCTIONELE EISEN | 15 |
| 5.1 COLLABORATORY..... | 17 |
| 5.2 ZOEKEN EN VINDEN..... | 17 |
| 5.3 ACHTERGRONDINFORMATIE..... | 18 |
| 5.4 PRESENTATIE | 18 |
| 5.5 ANALYSE | 18 |
| 5.6 REFEREREN | 19 |
| 5.7 PERSONALISATIE EN SAMENWERKING | 19 |
| 6 CONCLUSIES EN AANBEVELINGEN | 20 |
| BIJLAGEN | 22 |

Samenvatting

Van de digitale krantendatabank wordt een belangrijke impuls verwacht voor innovatief onderzoek in de geestes- en maatschappijwetenschappen. Wetenschappelijk onderzoekers vormen daarom een belangrijke doelgroep. Om er voor te zorgen dat de databank nieuw en innovatief onderzoek kan faciliteren is het van belang om goed aan te sluiten bij hun wensen. Dit rapport biedt een verkenning van de (groepen) onderzoekers die gebruik zullen maken van dagbladen bij hun onderzoek, van de onderzoeksvragen die zij stellen en van de eisen aan de krantendatabank die daaruit voortvloeien.

Het gaat daarbij om onderzoekers op het terrein van de geestes- en maatschappijwetenschappen. Het gebruik van ICT-methoden en -technieken neemt in deze disciplines toe. Toch zijn ze nog niet zozeer gemeengoed als in de meeste levenswetenschappen. De meeste projecten waarin nieuwe ICT-methoden en technieken worden toegepast komen voort uit technologisch onderzoek. De projecten die in het CATCH-programma worden uitgevoerd vormen hiervan een goed voorbeeld.¹

Dat ICT-methoden en -technieken nog geen gemeengoed zijn binnen de geestes- en maatschappijwetenschappen wordt gedeeltelijk verklaard door het feit dat onderzoekers in deze wetenschappen vaak gebruik maken van bronnenmateriaal dat nog niet digitaal beschikbaar is. Het is al een belangrijke stap voorwaarts als dat materiaal eenvoudig toegankelijk en bruikbaar wordt gemaakt. Het digitaliseren van bronnen vormt in de keten van innovatief onderzoek vaak de eerste stap en is dus belangrijk om nieuwe soorten onderzoek van de grond te krijgen. Natuurlijk zijn er ook onderzoekers die wel intensief gebruik maken van ICT-methoden en -technieken. In deze verkenning is getracht een balans te vinden tussen onderzoekers uit verschillende wetenschappelijke disciplines en tussen de onderzoekers wel en niet voorop lopen bij het gebruik van ICT-methoden en -technieken.

Wellicht speelt het kleine aandeel van ICT-methoden en -technieken een rol in de lage respons op het verzoek deel te nemen aan een interview of workshop voor deze verkenning. Het belang van de digitale krantendatabank wordt alom onderkend en men verwacht dat hierdoor nieuw en innovatief onderzoek mogelijk wordt. De meeste onderzoekers denken echter niet dat ze een bijdrage kunnen leveren aan het invullen van de eisen en wensen met betrekking tot een databank. Daardoor was het niet eenvoudig om geschikte kandidaten te vinden voor zowel de interviews als de workshop. Veelzeggend is in dit verband dat de benaderde archieven die kranten als bron aanbieden zich veel meer moeite getroostten om een antwoord op de gestelde vragen te vinden.

De verkenning laat zien dat kranten binnen een breed scala van disciplines in de geestes- en maatschappijwetenschappen een bron van informatie kunnen zijn (bijlage 1). In de meeste gevallen vormen ze echter maar één van de gebruikte bronnen. Gezien de grote verscheidenheid aan disciplines waar kranten als bron worden gebruikt wekt het geen verwondering dat er ook een grote diversiteit aan onderzoeksvragen bestaat. Toch zijn er wel overeenkomsten en categorieën te distilleren.

Een eerste categorie is onderzoek naar belangrijke thema's in de maatschappij. Kranten kunnen snel achtergrondinformatie geven over een onderzoeksonderwerp. Wanneer wordt er voor het eerst geschreven over nozems, bijvoorbeeld. Naast bronnen van inhoudelijke informatie vormen kranten vaak een afspiegeling van de maatschappij op een bepaald moment. Signatuur, verspreidingsgebied, aantal lezers en dergelijke vormen daarom een goede bron. Voor communicatie en mediawetenschappen vormt de krant als geheel een belangrijke bron. Als laatste vormt de krant als groot tekstcorpus een wezenlijke bron van onderzoek voor onderzoekers die zich bezighouden met taal en tekst. Zij zijn geïnteresseerd in de mogelijkheden die de krant als tekstcorpus biedt voor onderzoek.

¹ Continuous Access to Cultural Heritage (CATCH) <http://www.nwo.nl/catch> <geraadpleegd: 26-03-2008>

De eisen die onderzoekers stellen aan een digitale krantendatabank worden hier kort geschetst. Een basis-eis vloeit voort uit de behoefte bij onderzoekers om de juiste informatie te kunnen zoeken en vinden in een groot en divers corpus als de krantendatabank. Deze eis heeft bij hen dan ook de hoogste prioriteit. Men acht het van wezenlijk belang dat er op woordniveau gezocht kan worden. De kranten moeten dus *full text* ontsloten worden.

Daarnaast vinden onderzoekers het belangrijk dat de gevonden informatie snel geduid kan worden: gaat het om een artikel op de voorpagina, een opiniërend artikel, een ingezonden brief of andere rubrieken? De relatie tussen alle gegevens van kranten is van cruciaal belang voor het aanbieden van goede zoek- en analysemogelijkheden. Een ontologie waarin relaties aangebracht kunnen worden is hier van belang. Het is voor onderzoekers van wezenlijk belang dat deze informatie bij een artikel wordt opgeslagen en in zoekresultaten beschikbaar is.

Onderzoekers willen bovendien graag geïnformeerd worden over de betrouwbaarheid van hun zoekresultaten. De inhoud van de kranten wordt naar tekstbestanden omgezet in een automatisch OCR²-proces, dat veelal geen 100% correcte uitvoer op zal leveren. Men wil graag inzicht hebben in de mate van betrouwbaarheid van dat proces naar het gebruik van bepaalde kranten, periodes en dergelijke. Hulpmiddelen als opzoeklijsten, waarbij gezocht kan worden op persoonsnamen of geografische namen, worden bijzonder bruikbaar geacht. Daarnaast willen de onderzoekers gebruik kunnen maken van geavanceerde zoekopties zoals stemming en *fuzzy search*. Ook wil men de krantendatabank graag gekoppeld zien aan externe bronnen, zoals geografische en genealogische bestanden.

Een zoekopdracht binnen de krantendatabank kan veel informatie opleveren. Onderzoekers willen die graag zo geordend en overzichtelijk mogelijk gepresenteerd krijgen. Daarom is het van belang dat de informatie kan sorteren per krant, periode, rubriek en dergelijke kan worden gepresenteerd, terwijl ook combinaties van bijvoorbeeld periode, rubriek en krant mogelijk moeten zijn. Ook wordt waarde gehecht aan de mogelijkheid om zowel het afzonderlijke artikel als de gehele pagina en krant te kunnen inzien. Naar gelang een onderzoeker verdere analyse op het materiaal wil toepassen zal hij of zij er belang aan hechten dat de artikelen in een aantal formaten beschikbaar zijn: als PDF, als *full text* en in de vorm van xml bestanden.

Onderzoekers geven aan dat de mogelijkheid om annotaties tijdens het onderzoeksproces aan te brengen dat proces goed kan ondersteunen. Deze annotaties moeten op een gepersonaliseerde wijze toegevoegd en achteraf door de onderzoeker gedownload kunnen worden. Ook wordt het wenselijk gevonden om op een eenduidige wijze te kunnen refereren naar artikelen op basis van unieke identifiers. Die identifiers kunnen ook door de onderzoeker gebruikt worden om relaties aan te brengen tussen artikelen om zodoende afhankelijkheden in kaart te brengen.

Onderzoekers die de krantendatabank gebruiken als tekstcorpus willen graag grote hoeveelheden artikelen kunnen selecteren en als xml bestanden downloaden. Deze bestanden worden dan gebruikt om analysetools op los te laten teneinde die verder te ontwikkelen. Een andere mogelijkheid die geopperd wordt is dat de krantendatabank dusdanig is ingericht dat analysetools direct op het corpus kunnen worden gebruikt.

Er is uitdrukkelijk gevraagd naar de behoefte aan een zogenaamde collaborative functionaliteit. Die bleek bij het merendeel van de onderzoekers niet aanwezig. Moet er worden samengewerkt aan hetzelfde onderzoek, dan heeft iedereen zijn eigen deel en hoeven de gevonden artikelen niet gedeeld te worden. Een aantal onderzoekers ziet wel iets in het opbouwen van een persoonlijke bookmarkcollectie die gedeeld kan worden met anderen. Hierdoor ontstaat dan een soort peer-review systeem van artikelen. Het merendeel van de onderzoekers acht dit niet direct wenselijk gezien het feit dat de krantendatabank slechts een van de vele bronnen is die bij het onderzoek gebruikt worden.

Voor veel onderzoekers is informatie over de krant zelf van belang: het verspreidingsgebied, het aantal lezers, de signatuur van de krant en biografische gegevens van de journalisten. Het koppelen van deze pershistorische informatie aan de krantendatabank heeft een grote meerwaarde om de context van de gevonden informatie te kunnen duiden.

² Optical Character Recognition

1 Opdracht

1.1 Aanleiding

In 2006 is het projectvoorstel 'Digital Databank for Newspapers' van de Koninklijke Bibliotheek (KB) door het innovatieplatform uitgekozen om in in aanmerking te komen voor financiering. Het project behelst het ontwikkelen en exploiteren van een digitale databank voor dagbladen. Het belangrijkste doel is om een impuls te geven aan het opzetten van een grootschalige data-infrastructuur, die op zijn beurt nieuw en innovatief onderzoek voor de humaniora en sociale wetenschappen stimuleert en faciliteert. Wetenschappelijk onderzoekers zijn dus de voornaamste doelgroep voor de databank.

Het project, dat wordt uitgevoerd door de hoofdafdeling Research & Development van de KB, heeft een looptijd van vijf jaar en beslaat de periode 1618-1995. De inhoud van de databank zal bestaan uit ongeveer acht miljoen pagina's van nationale, regionale, lokale en koloniale dagbladen.

Het succes van het project sterk afhangen van het gebruik van de databank door onderzoekers. Daarom is het van belang dat de inhoud van de krantendatabank de onderzoeksvragen van wetenschappelijke onderzoekers ondersteunt. Daarnaast moet de manier waarop de kranten worden ontsloten nauw aansluiten bij de methoden en technieken die onderzoekers gebruiken en deze op een adequate wijze faciliteren.

Om op deze gebieden zo goed mogelijk tegemoet te komen aan de wensen van de onderzoekers is het van belang deze doelgroep in een vroeg stadium bij het project te betrekken. Zowel bij het selecteren van de inhoud als ten aanzien van de wijze van ontsluiting moeten keuzes worden gemaakt en prioriteiten gesteld, waarvoor voldoende draagvlak bestaat in het wetenschappelijk veld.

In het projectvoorstel wordt aan de KNAW de rol toegedacht om onderzoekers te betrekken bij de wijze van ontsluiting en toegang tot de digitale dagblad collecties. Deze verkenning levert daaraan een bijdrage.

1.2 Doel en opzet van het onderzoek

Om de onderzoekers in de humaniora en sociale wetenschappen adequaat te kunnen bedienen is inzicht vereist in:

- de onderzoekers of groepen onderzoekers die gebruik maken van dagbladen als bron;
- de onderzoeksvragen zij stellen;
- de eisen aan de krantendatabank die daaruit voortvloeien;
-

De opdracht aan DANS is om dit in kaart te brengen. Daartoe heeft DANS de onderstaande activiteiten verricht:

1. Inventariseren in welke wetenschappelijke disciplines kranten als bron gebruikt worden en welke onderzoekers of groepen onderzoekers daarbij betrokken zijn.
2. In kaart brengen welke functionaliteit momenteel in digitale krantendatabanken aangeboden wordt.
3. Inventariseren met welke (inter)nationale ontwikkelingen rond digitale krantendatabanken rekening moet worden gehouden.
4. Inventariseren welke onderzoeksvragen er door de betrokken onderzoekers gesteld worden.
5. De behoeften van onderzoekers inzichtelijk maken met betrekking tot de functionaliteit van de digitale krantendatabank.
6. Prioriteiten stellen met betrekking tot de functionaliteit zoals toegankelijkheid, zoekfunctionaliteit, analysegereedenschappen en mogelijkheden tot samenwerking.
7. De functionele eisen rapporteren waaraan de digitale krantendatabank dient te voldoen.

1.3 Afbakening en bereik

Het bereik van de inventarisatie is als volgt afgebakend

- De doelgroep bestaat uit wetenschappelijk onderzoekers in de sociale- en geesteswetenschappen;
- Bij de kranten gaat het om Nederlandse nationale, regionale, lokale en koloniale dagbladen;
- De gestelde vragen moeten met beperkte middelen worden beantwoord. De uitkomsten in dit rapport zijn niet uitputtend maar wel representatief van aard.
- DANS levert een bijdrage aan de besluitvorming over keuzes die in het project 'Digital Databank for Newspapers' gemaakt moeten worden met betrekking tot de zoekfunctionaliteit, de analyse *tools* en *collaboratories*, maar is niet verantwoordelijk voor die keuzes.

1.4 Beoogde resultaten

Het resultaat van het project bestaat uit één rapport waarin het volgende wordt beschreven.

1. De (groepen) onderzoekers die kranten gebruiken als bron bij hun onderzoek en de onderzoeksvragen die zij stellen.
2. Een beknopte inventarisatie van relevante (inter)nationale ontwikkelingen.
3. De gewenste functionele eisen die vanuit de wetenschap aan de krantendatabank gesteld worden.

2 Verloop van de verkenning

2.1 Selectie van onderzoekers

Om inzicht te krijgen in de behoeften van onderzoekers waar het de functionaliteit van de digitale krantendatabank aangaat, wordt een aantal onderzoekers geselecteerd om te worden geïnterviewd en voor deelname aan een workshop. De selectie moet zoveel mogelijk de wetenschappelijke disciplines vertegenwoordigen waar onderzoekers gebruik maken van kranten als bron.

Met behulp van de Nederlandse Onderzoeksdatabank (NOD)³ is een overzicht gemaakt van de wetenschappelijke disciplines die onder de sociale- en geesteswetenschappen vallen. Om vervolgens binnen zoveel mogelijk disciplines onderzoekers te vinden die kranten als bron gebruiken zijn de volgende stappen gedaan:

- Aan de hand van de projectdocumentatie en andere relevante informatie is gekeken welke onderzoekers op enige wijze bij het project betrokken zijn;
- Bij vijf aanbieders van digitale kranten is nagegaan welke onderzoekers er gebruik van maken;
- Er is onderzocht welke onderzoekers verwijzen naar kranten in hun wetenschappelijke publicaties.
- Er is aan enkele onderzoekers in verschillende disciplines gevraagd of men binnen deze disciplines kranten gebruikt voor onderzoek en welke onderzoekers dat zijn.

Het resultaat was een zogenaamde *longlist* van onderzoekers die benaderd kunnen worden voor de interviews of de workshop.

2.2 Technische mogelijkheden korte termijn

Door middel van een *quicksan* van leveranciers van software voor digitale krantendatabanken is de huidige functionaliteit in kaart gebracht. De KB heeft dit werkpakket uitgevoerd. DANS is bij deze gesprekken aanwezig geweest. De *quicksan* levert een overzicht op van de functionaliteit die op dit moment door leveranciers aangeboden wordt.

2.3 Technische mogelijkheden lange termijn

Door middel van een beknopte desk research is in korte tijd geïnventariseerd met welke (inter)nationale ontwikkelingen rekening moet worden gehouden. De desk research geeft inzicht in beschikbare functionaliteit die op de middellange termijn gebruikt zou kunnen worden. Aan de hand van de Nationale Inventarisatie Krantendigitalisering⁴ (Digitaal Erfgoed Nederland) is een inventarisatie gemaakt van de manieren waarop digitale kranten momenteel in Nederland ontsloten worden. Met behulp van de website van de International Coalition on Newspapers (ICON)⁵ is vervolgens de internationale stand van zaken in beeld gebracht. Om een indruk te krijgen van de nationale en internationale trends in het ontsluiten van grote hoeveelheden tekstuele gegevens, is gekeken naar relevante projecten binnen het door NWO gefinancierde CATCH-programma en naar projecten die gesubsidieerd worden door de kaderprogramma's van

³ Nederlandse Onderzoeks Databank (NOD) <http://www.onderzoekinformatie.nl/nl/oi/nod/>
<geraadpleegd: 26-03-2008>

⁴ Nationale Inventarisatie Krantendigitalisering
http://matrix.den.nl/matrix.aspx?matrixid=krantendigitalisering&view=Digitaal_Erfgoed
<geraadpleegd: 29-03-2008>

⁵ International Coalition on Newspapers (ICON) <http://icon.crl.edu/digitization.htm> <geraadpleegd: 29-03-2008>

de Europese Commissie⁶. Aansluitend is nog een aantal onderzoekprojecten geëvalueerd op het multidisciplinaire terrein van taal, tekst en informatica. Het overzicht van de mogelijke gebruikersfunctionaliteit gaf informatie aan de onderzoekers die geïnterviewd zijn en diende als input voor discussie tijdens de workshop.

2.4 Onderzoeksvragen en gewenste functionaliteit

Op basis van de inventarisatie zijn vijf onderzoekers geselecteerd die kranten als belangrijke bron voor hun onderzoek gebruiken. Ze vertegenwoordigen de disciplines Nederlands, geschiedenis, wetenschapsfilosofie, algemene sociale wetenschappen en taaltechnologie. Hun onderzoeksvragen en hun behoeften op het gebied van de functionaliteit van de databank zijn met behulp van *face to face* interviews in kaart gebracht. Ook is een aantal telefonische interviews gehouden bij de aanbieders van digitale kranten(-archieven) om een beeld te krijgen van de onderzoekers die er gebruik van maken en de disciplines waarin zij werken.

2.5 Workshop

Op basis van de interviews en het verworven inzicht in de technische stand van zaken is een workshop georganiseerd waarin met vijf onderzoekers de behoefte aan functionaliteit met betrekking tot toegankelijkheid, zoeken, analysegereedschappen en mogelijkheden tot samenwerking is besproken. Zij waren werkzaam in de volgende wetenschappelijke disciplines: journalistiek, archivaliek, massacommunicatie, taal en letterkunde, politicologie, sociologie en geschiedenis.

2.6 Terugkoppeling

De resultaten zijn verwerkt in dit rapport en worden teruggekoppeld naar de betrokken onderzoekers.

⁶ European Research Framework Programme. http://cordis.europa.eu/fp7/home_en.html
<geraadpleegd: 28-03-2008>

3 Onderzoekers en hun onderzoeksvragen

Om te achterhalen welke onderzoekers binnen de humaniora en sociale wetenschappen gebruik maken van kranten als bron, zijn enkele archieven benaderd die digitale landelijke dagbladen aanbieden. Die werden geselecteerd op basis van het projectenoverzicht van Digitaal Erfgoed Nederland (DEN). Aan deze aanbieders zijn telefonisch de volgende vragen gesteld:

- Zijn er wetenschappelijke onderzoekers die gebruik maken van de aangeboden kranten voor hun onderzoek?
- Uit welke wetenschappelijke discipline of welk onderzoeksgebied zijn zij afkomstig?
- Welke onderzoeksvraag willen zij beantwoorden met de informatie van de kranten?
- *Eventueel*: Welke onderzoeksmethode gebruiken zij hiervoor?

In gesprek met de aanbieders van gedigitaliseerde dagbladen⁷ werd duidelijk dat er bij hen geen overzicht bestaat van de gebruikers. Dit komt grotendeels door die digitalisering zelf: de gebruikers kunnen gemakkelijk zelf zoeken via het web, waardoor er weinig tot geen contact meer is tussen de aanbieders en de onderzoekers.

De telefonisch benaderde medewerkers van archieven hebben wel een globaal idee over wat voor soort gebruikers de digitale kranten raadplegen. Het betreft voornamelijk liefhebbers, hobbyisten en genealogen en ook wel studenten en journalisten. Academische onderzoekers maken sporadisch gebruik van de huidige digitale kranten. Met name genoemd worden historici, archeologen en architecten.

Om het beeld nog iets aan te scherpen is met behulp van SciSearch⁸ gezocht naar publicaties met verwijzingen naar Volkskrant, Telegraaf, NRC Handelsblad, Utrechts Nieuwsblad en het Algemeen Dagblad. De onderzoekers die meer dan één krant als bron voor hun publicatie bleken te hebben gebruikt zijn geselecteerd voor de longlist van potentiële kandidaten voor een interview. Hierbij is rekening gehouden met een zo goed mogelijke spreiding over de disciplines binnen de sociale- en geesteswetenschappen.

Uit de inventarisatie blijkt dat binnen zeer veel wetenschappelijke disciplines kranten als bron gebruikt worden (bijlage 1). Het is moeilijk te beoordelen welke andere disciplines hier in de toekomst nog aan toegevoegd kunnen worden.

Hoewel deze uitkomsten maar een eerste indicatie geven, kan voorzichtig de conclusie worden getrokken dat in potentie bijna alle wetenschappelijke disciplines in meer of mindere mate van deze bron gebruik maken.

Aan onderzoekers is vervolgens expliciet gevraagd welke onderzoeksvragen zij op dit moment stellen waarbij kranten een wezenlijke bron van informatie bevatten. De verscheidenheid blijkt groot. Hieronder een korte impressie:

- De commercialisering van de universiteiten.
- Democratie
- Alternatieve werkverbanden, alternatieve geldstelsels
- Het verschil tussen publiek en privaat onderzoek over *Human genome* in 2003
- Hoe functioneert het basisinkomen in Amerika (bepaalde periode).
- Hoe verlopen lokale referenda?
- Het TBS-systeem, de ontsnappingen in het bijzonder en de reactie van de politici hierop.
- De ‘gabbercultuur’

⁷ Er is gesproken met vijf archieven die digitale kranten online toegankelijk hebben gemaakt.

⁸ SciSearch®: A Cited Reference Science Database is an international, multidisciplinary index to the literature of science, technology, biomedicine, and related disciplines produced by Thomson (ISI®). SciSearch contains all of the records published in the *Science Citation Index*® (SCI®), plus additional records in engineering technology, physical sciences, agriculture, biology, environmental sciences, clinical medicine, and the life sciences. SciSearch indexes all significant items (articles, review papers, meeting abstracts, letters, editorials, book reviews, correction notices, etc.) from more than 6,100 international scientific and technical journals.

- De thuiszorg in Nederland
- De geschiedenis van de wetenschap
- Op welk moment dringen begrippen vanuit de wetenschap door in de populaire wetenschap?
- Hoe heeft patent zich de laatste tweehonderd jaar ontwikkeld?
- Hoe ontwikkelt de taal zich op een bepaald moment?
- Kranten zijn als tekstcorpus gebruikt bij het ontwikkelen van statistische taalmodellen, die informatie bevatten over de meest voorkomende opeenvolging van woorden in een taal. Dit vormt de basis voor het automatisch transcriberen van gesproken taal.

De onderzoeksvragen blijken heel verschillend te zijn. De krant wordt als bron gebruikt om informatie te verzamelen, om bepaalde thema's en onderwerpen gedurende een bepaalde periode te volgen of om de weergave van feiten over personen, gebeurtenissen en onderwerpen te onderzoeken. Meerdere onderzoekers benadrukken dat kranten zeer zeker niet hun enige bron zijn. Daarnaast verschilt de rol die kranten binnen een onderzoek vervullen naar gelang de onderzoeker. Voor sommigen helpt de krant bij het inventariseren, anderen maken er gebruik van voor een kwalitatieve verkenning van een thema of het volgen van een onderwerp in de tijd, om zo ideeën op het spoor te komen.

Om een indruk te krijgen van de mogelijkheden van de digitale krantendatabank als bron is aan onderzoekers ook gevraagd welke vragen een rol kunnen gaan spelen als de kranten digitaal tot hun beschikking staan. Een impressie:

- Ontwikkeling van woordbetekenis door de tijd heen;
- Frequenties van woordcombinaties en de veranderingen daarin;
- Taalmodellen voor voorbije periode;
- Onderzoek naar intertekstualiteit. Hoe staan verschillende teksten met elkaar in verband en hoe hebben zij invloed op elkaar? Of: wie schrijft er het eerst over een bepaald onderwerp en wie pikt het verder op; kortom de beïnvloedingsrelaties zichtbaar maken;
- Taalmodellen willen construeren voor oudere fasen van het Nederlands ('daarvoor is dit natuurlijk interessant');
- Vergelijkende vragen over een onderwerp over de tijd (de nozems vs de gabbers of de gabbers van toen en nu);
- Internationaal vergelijkend onderzoek (jeugdculturen in Engeland en Nederland);
- Meer onderzoek naar/vanuit regionale dagbladen.

Hoewel het om een heterogene verzameling vragen gaat uit verschillende disciplines, kunnen met enige creativiteit wel clusters worden onderscheiden. In de eerste plaats gebruiken onderzoekers kranten vaak om achtergrondinformatie te verzamelen, zich te oriënteren op hun onderwerp of ter inspiratie. Dus: exploreren van het onderwerp, inventariseren en oriënteren om de onderzoeksvraag verder richting te kunnen geven. In de tweede plaats kunnen kranten door middel van kwalitatieve of kwantitatieve verkenning inzicht geven in maatschappelijke thema's, personen of gebeurtenissen. Daarnaast worden kranten veel als bron gebruikt voor het onderzoek naar personen, gebeurtenissen of thema's. De feitelijke verslaglegging wordt dan gebruikt om inhoudelijk onderzoek te onderbouwen, bijvoorbeeld over een historische gebeurtenis, de berichtgeving erover of de impact ervan.

Door de continuïteit van de berichtgeving in kranten vormen zij, in de vierde plaats, een belangrijke bron voor longitudinaal onderzoek. Onderwerpen kunnen in de tijd gevolgd worden en trends en breuken kunnen daardoor onderzocht worden. In de vijfde plaats spelen kranten een belangrijke rol bij vergelijkend onderzoek, bijvoorbeeld tussen de nozem cultuur en de gabber cultuur als het over sociaal maatschappelijke thema's gaat. En tenslotte worden kranten gebruikt om onderzoeksvragen te kunnen beantwoorden die longitudinaal en vergelijkend onderzoek met elkaar combineren.

Behalve als inhoudelijke bron worden kranten ook gebruikt voor onderzoek op het gebied van de taal en letterkunde, bijvoorbeeld voor onderzoek naar intertekstualiteit, relaties tussen teksten, het bouwen van statistische taalmodellen en fasen van een taal door de tijd heen: de krant als

tekstcorpus. Aansluitend op deze laatste vorm van onderzoek wordt de krant ook vaak gebruikt om nieuwe methoden en technieken zoals *text mining* op hun bruik- en betrouwbaarheid te controleren.

Als voordeel van een digitale krantendatabank ten opzichte van een papieren krantenarchief kan voorzichtig geconcludeerd worden dat onderzoekers hierdoor meer en betere mogelijkheden voor het verrichten van longitudinaal en vergelijkend onderzoek verwachten.

Aan de onderzoekers is tenslotte gevraagd of er door hen interdisciplinair onderzoek wordt gedaan met kranten als bron en of zij verwachten dat samenwerking tussen verschillende wetenschappelijke disciplines bevorderd wordt door de digitale krantendatabank. Het blijkt dat er af en toe interdisciplinair wordt samengewerkt in onderzoek met behulp van kranten. Disciplines die worden genoemd zijn: combinaties van informatici en mensen met een taalwetenschappelijke achtergrond, geschiedenis, politicologie, psychologie, onderwijskundigen, sociale wetenschappen, economie en methoden & technieken in verband met de methoden van onderzoek.

Niet alle wetenschappers geven aan interdisciplinair te werken. Dit zou komen doordat verschillende wetenschapsdisciplines zelf barrières voor deze samenwerking opwerpen en door de verschillen in onderzoeksparadigma's en definities van begrippen. De personen die aangeven dat er op dit moment al niet veel interdisciplinair wordt gewerkt verwachten hierin ook geen veranderingen in de toekomst. Zij verwachten daardoor niet van de krantendatabank dat deze samenwerking tussen onderzoekers gefaciliteerd. De andere wetenschappers denken vooral aan combinaties met disciplines waarbij een historische invalshoek een rol speelt. Sommige onderzoekers geven aan geïnteresseerd te zijn in het aanleggen van een persoonlijke *bookmark* collectie die met collega wetenschappers gedeeld kan worden.

4 Gebruikersfunctionaliteit

4.1 Korte termijn

Binnen het aanbod van gedigitaliseerde kranten op het internet is bekeken welke functionaliteit voor onderzoek beschikbaar is. De gevonden functionaliteit kan grofweg onderverdeeld worden in drie categorieën: zoekopties, zoekmechanismen en presentatie. Onder zoekopties worden de zaken genoemd waarop gezocht kan worden, onder zoekmechanismen vallen de technieken hoe er gezocht kan worden en onder presentatie valt hoe het resultaat bekeken kan worden. Een overzicht van deze functionaliteit bevindt zich in de bijlage.

Uit deze inventarisatie blijkt dat de hoeveelheid zoekopties afhankelijk is van de fijnmazigheid waarmee de verschillende onderdelen van een krant toegankelijk zijn gemaakt. Hoe meer onderscheid er wordt gemaakt tussen de verschillende onderdelen van een krant des te verfijnder kan er worden gezocht. Om een indruk te krijgen van de beschikbare elementen van een krant en de relaties hiertussen is een fysieke krant theoretisch ontleed en zijn deze elementen en relaties beschreven middels een ontologie. Het resultaat hiervan bevindt zich in de bijlage.

4.2 Middellange termijn

Binnen het NWO-programma CATCH en het Europese *Framework Programme* is gekeken naar onderzoeken die zich bezig houden met toegankelijkheid van grote tekstcorpora en naar onderzoeken waar de focus ligt op de analyse van tekstueel en visueel materiaal. Van deze projecten is kort bekeken wat ze doen en wat hun resultaten zouden kunnen bijdragen aan de digitale krantendatabank (zie bijlage).

De geraadpleegde onderzoeksprojecten houden zich met uiteenlopende onderwerpen bezig en hebben vaak onderlinge relaties. De technieken waarmee ze worden uitgevoerd, kunnen in drie groepen worden onderverdeeld:

1. analyse (Tekstanalyse, Beeldanalyse, Geluidanalyse)
2. annotatie (Ontologie, Tijdschaal, Personalisatie, Aanbevelingen)
3. integratie (Semantisch Web, *Collaborations*)

De relatie tussen deze drie groepen is steeds dezelfde: analyse geeft informatie over de data die vervolgens via annotatie kunnen worden vastgelegd, waarna met behulp van deze informatie de data kunnen worden geïntegreerd met andere data. Op middellange termijn kunnen de methoden en technieken uit deze onderzoeksprojecten toegevoegde waarde hebben voor de toegankelijkheid en analysemogelijkheden van de digitale krantendatabank.

In de informatica wordt veel onderzoek gedaan naar *Information Retrieval*. Binnen deze discipline is in het kader van de desk research gekeken naar relevante mogelijkheden op het terrein van *Text Mining* voor het ontsluiten van kranten. *Text mining* wordt in het algemeen gedefinieerd als het proces waarmee kwalitatief hoogwaardige informatie uit grote hoeveelheden veelal ongestructureerde tekst wordt gehaald. Enkele methoden en technieken die bij *text mining* gebruikt worden kunnen relevant zijn voor het ontsluiten van kranten. Te denken valt hierbij aan:

- *text categorization*: het handmatig of automatisch toekennen van categorieën aan teksten gebaseerd op de inhoud van deze teksten;
- *text clustering*: het automatisch groeperen van artikelen op basis van gemeenschappelijk woorden;
- *concept/entity extraction*: het extraheren van woorden of zinnen uit een tekst die zo goed mogelijk de inhoud van die tekst weergeven;
- *taxonomies*: het aanbrenge van classificaties en relaties;
- *sentiment analysis*: het determineren van de attitude van een schrijver ten opzichte van een bepaald onderwerp;
- *document summarization*: het automatisch maken van een samenvatting van een tekst;

- *named entity*: het aanbrengen van voorgedefinieerde categorieën in teksten met betrekking tot bijvoorbeeld personen, organisaties, locaties.

De methoden en technieken die bij text mining worden ontwikkeld kunnen een bijdrage leveren aan een verbeterde bruikbaarheid van de krantendatabank. De technieken lopen sterk uiteen. Zij kunnen op de achtergrond analyses uitvoeren en de krantendatabank verrijken, maar bij onderzoekers bestaat ook de wens hun eigen tools met de databank te laten communiceren. Omdat zulke communicatie specialistisch van aard is zou deze via een speciale interface kunnen verlopen.

Uit de CATCH-onderzoeksprojecten en uit de analyse van het aanbod van digitale kranten zijn dus functionaliteiten gedestilleerd die voor het ontsluiten van kranten relevant is. Om dit helder weer te geven worden deze functionaliteiten in het overzicht hieronder opgedeeld in *toegankelijkheid*, *presentatie* en *gebruik*. De functionaliteiten zijn vervolgens gekoppeld aan mogelijke methoden en technieken van wetenschappelijk onderzoek. Het overzicht vormt zo een palet aan mogelijke functionaliteiten, welke in de digitale krantendatabank verwezenlijkt zouden kunnen worden.

| Functionaliteit | Techniek |
|---|---|
| <i>Toegang</i> | |
| <ul style="list-style-type: none"> • Zoeken <ul style="list-style-type: none"> ○ Meerdere collecties ○ Granulariteit ○ Chronologisch ○ Geografisch ○ Semantisch • Refereren • Bladeren • Interoperabiliteit <ul style="list-style-type: none"> ○ Collecties ○ Opslaan / exporteren | <ul style="list-style-type: none"> querytaal standaardisatie van querytaal en metadata / vertalingsmechanisme voor queries en metadata uitgebreide krantontologie/fuzzy search zoeken via een tijdbalk zoeken via een kaart (GIS) Semantisch Web/ontologie/thesaurus provide persistent identifiers/enable deeplinking /provide 'cite this' info bladeren kan via groeperingen, directorystructuren, tagclouds, etc.d standaardisatie van metadata standaardisatie van bestandsformaten |
| <i>Presentatie</i> | |
| <ul style="list-style-type: none"> • Artikelen <ul style="list-style-type: none"> ○ Granulariteit • Zoekresultaat <ul style="list-style-type: none"> ○ Groeperen ○ Geografisch ○ Chronologisch | <ul style="list-style-type: none"> tekst, grafisch, voorlezen, highlighten van gevonden zoektekst highlighting woorden of artikel textclustering, 'group by {property}' presentatie op kaart (GIS) presentatie op tijdbalk |
| <i>Gebruik</i> | |
| <ul style="list-style-type: none"> • Koppelen • Personaliseren • Waarderen • Herkenning <ul style="list-style-type: none"> ○ Uit beeld ○ Uit tekst • Vergelijken • Categoriseren • Samenwerken | <ul style="list-style-type: none"> provide persistent identifier/ontology/kunstmatige intelligentie verzorg annotatie- en bookmarkmogelijkheden sentiment analyse/kunstmatige intelligentie beeldherkenning/gezichtsherkenning/kunstmatige intelligentie information extraction/entity matching/kunstmatige intelligentie waarderen/kunstmatige intelligentie text categorization forum, web 2.0, collaboration |

- Integriteit bewaren query en resultaat, versiebeheer, fout-detectie/correctie
- 'Rekenintensief' onderzoek opslag- en rekencapaciteit van GRID-technologie

5 Functionele eisen

Uit de telefonische enquête komt naar voren dat aanbieders van digitale kranten (archieven) een globaal beeld hebben van hetgeen gebruikers belangrijke functionaliteiten vinden. Men wil graag *full text* kunnen zoeken. Foto's zijn bij voorkeur gekoppeld aan onderwerpen en kunnen besteld of uitgeprint worden. Men wil bij voorkeur fonetisch kunnen zoeken en genealogische databanken kunnen koppelen aan de digitale kranten. De zoekopdrachten wil men graag kunnen verkleinen door middel van herhaald zoeken in de gevonden resultaten en men wil graag trefwoorden kunnen koppelen aan periode en locatie.

Uit deze opsomming komt duidelijk naar voren dat de gebruikersgroep hier voor het merendeel bestaat uit genealogen en andere geïnteresseerde gebruikers en dat men in mindere mate zicht heeft op de behoeften van wetenschappelijke onderzoekers met betrekking tot gewenste functionaliteiten.

Voorafgaande aan de *face-to-face* interviews hebben twee inventariserende gesprekken plaatsgevonden, om een eerste indruk te krijgen van de mening van de onderzoekers en om de opzet van de uiteindelijke interviews te bepalen. Daarna volgden vijf interviews met onderzoekers met als disciplines: Nederlands, geschiedenis, wetenschapsfilosofie, algemene sociale wetenschappen en taaltechnologie. Tijdens deze interviews is behalve naar de onderzoeksvragen die men wil stellen ook naar de wensen gevraagd met betrekking tot het raadplegen van de dagbladen.

De mogelijkheid om materiaal van de Digitale Databank te *downloaden* wordt door de geïnterviewden aangemerkt als zeer belangrijk. De toegevoegde waarde is dat ze het materiaal in eigen analysegereedschap kunnen importeren, kunnen printen en verdere bewerkingen kunnen uitvoeren op de eigen computer. De manier waarop er moet worden gedownload - in bulk of per artikel - verschilt echter per onderzoeker en per onderzoeksvraag.

Voor de auteursrechten geldt dat de auteurs die zo geregeld willen zien dat zij zo min mogelijk beperkt worden in hun onderzoek. Uit de interviews blijkt dat wetenschappers niet gebonden willen zijn aan de functionaliteiten die de Databank biedt en zal gaan bieden. Eén van de geïnterviewden wil benadrukken dat het een gemiste kans zou zijn als het systeem op welke manier dan ook gesloten zou zijn.

Annotaties kunnen maken in de Digitale Databank blijkt voor de geïnterviewden een belangrijke functionaliteit. Deze moeten vervolgens wel te downloaden zijn. Men wil voetnoten en bibliografieën kunnen maken, *highlighten* en een persoonlijke *bookmark* collectie kunnen aanleggen. Eén van de geïnterviewden oppert dat annotaties achteraf, *social tagging*, door middel van *recommender* technologie wellicht een interessante extra mogelijkheid is.

Bij het aan elkaar koppelen van materiaal uit de Digitale Databank wordt door de geïnterviewden de kanttekening geplaatst dat het moeilijk is de betrouwbaarheid hiervan te garanderen. Eén van de geïnterviewden merkt op dat door het koppelen van materiaal informatie verloren gaat: 'weggooien van informatie kan altijd nog'. Aanverwante ideeën van de geïnterviewden bij 'koppelen' zijn onder andere: gerelateerde artikelen weergeven, zoals lezersreacties, vervolg artikelen in latere nummers en achtergrondartikelen; het koppelen aan andere databanken wellicht met andere media; websites weergeven die in verband staan met de zoekopdracht.

Uit de interviews blijkt dat wat betreft zoekfunctionaliteiten men tot in het oneindige de zoekopdracht wil kunnen specificeren. Daarbij moet men kunnen aangeven hoe bepaalde zoektermen zich tot elkaar verhouden (bijv. de plaats van twee woorden ten opzichte van elkaar in een zin). Er wordt voorgesteld dat naast de zoekresultaten kwantitatieve gegevens kunnen worden weergegeven, zoals hoe vaak een bepaalde term voorkomt in de databank.

Vanuit het projectplan van de KB is de geïnterviewden een aantal zoekfunctionaliteiten voorgelegd (zie de bijlage). Daarnaast worden de aanvullende ideeën van de geïnterviewden puntsgewijs gepresenteerd.

Ideeën van geïnterviewden

Vinden

- Aantal abonnees in een bepaalde periode
- Woonplaats abonnees in een bepaalde periode
- Per editie in periode
- Gerelateerde artikelen
- Artikelen als reactie op
- Ingezonden brieven
- Artikelen die op andere pagina verder gaan
- Automatisch toekennen onderwerpcategorie
- Hiërarchie van onderwerpen
- Ook buiten categorieën om zoeken
- Performance --> snelheid is belangrijk(st)
- Beperkingen aan kunnen geven
- Per krant
- Over meerdere kranten
- Alternatieve zoekterm op basis van eigen zoekterm bieden
- Met steekwoord ook foto's kunnen vinden
- Nieuws versus achtergrond artikelen
- Titelwoord versus artikelwoord

Selecteren

- Op basis van kwaliteit van de tekst (OCR)
 - onderzoeker kent schaalscore toe van goed herkende ocr-uitvoer
 - foutenpercentage per jaargang
- Op basis van relevantie
 - tijd die onderzoeker op link doorbrengt

Analyseren

- Het aantal maal dat een bepaalde combinatie van woorden voorkomt
 - Werkwoord; zelfstandig naamwoord; plaats in de zin
- Niet gebonden aan functionaliteit van de Digitale Databank voor Dagbladen
 - alles kunnen downloaden, eigen tools
 - gesloten systeem is gemiste kans
 - uploaden en delen van eigen tools
 - netwerkanalyse, personen samen genoemd
- Krant als inventarisatie op onderwerp
- Altijd inzicht in analysetechniek houden
- Onderzoeker idee van totale controle
- (Her)categoriseren faciliteren
- Lezers review van artikel op basis van:
 - kwaliteit analyses door auteur
 - kwaliteit / hoeveelheid info
 - in hoeverre overzichtsartikel?
 - in hoeverre is artikel vormend idee voor onderzoeker
- Zien hoe vaak zoekterm wordt gebruikt

Bewerken

- Annotaties
- Highlighten
- Kunnen printen

Presenteren

- Trefferlijst zichtbaar naast artikel
- Artikel als thumbnail
- Tabbladen
 - Essentie op eerste blad
 - Extra's op bladen daar achter

Bewaren

- Persoonlijke bookmark collectie

Terugvinden

- Persoonlijke bookmark collectie
- Iedere pagina unieke url

5.1 Collaboratory

De onderzoekers blijken van mening dat de keus voor een onderzoeksvraag en een eventueel samenwerkingsverband aan de onderzoeker is en dat de krantendatabank hiervoor alleen de gedigitaliseerde kranten hoeft te leveren: verdere samenwerking hoeft een krantendatabank niet te faciliteren.

De interviews gaven een goed beeld van de wensen van onderzoekers met betrekking tot de krantendatabank. De workshop bood daarenboven de gelegenheid om nader op bepaalde wensen en problematiek in te gaan. Hieronder volgt een samenvatting van punten van belang. Het beeld dat uit de interviews naar voren kwam werd ook in de workshop bevestigd. Een kleine groep onderzoekers weet precies wat ze wil en is zich bewust van de mogelijkheden en onmogelijkheden die nieuwe technieken voor innovatief onderzoek meebrengen. Het merendeel van de onderzoekers is tevreden indien de kranten zonder teveel drempels raadpleegbaar zijn. De nadruk ligt in eerste instantie op het zoeken en vinden van relevante informatie. Daarnaast vindt men de wijze van presentatie van belang. Voor het analyseren van de gevonden informatie wil men graag de mogelijkheden hebben om het gevonden materiaal te kunnen downloaden zodat men er eigen analysetools op los kan laten.

5.2 Zoeken en vinden

Om goed te kunnen zoeken, vinden en analyseren moeten de kranten volledig op woordniveau doorzoekbaar zijn. Er wordt veel belang gehecht aan het zo correct mogelijk uitvoeren van het OCR-proces. Het streven gaat uit naar 100% correcte ocr-uitvoer. Voor onderzoekers is het van belang om een indicatie te hebben van de betrouwbaarheid van de gevonden gegevens, bij alle vormen van onderzoek die men op de krantendatabank zou willen toepassen: kwantitatief, kwalitatief thematisch, longitudinaal, linguïstisch. Alle aanwezigen zijn goed op de hoogte van het gegeven dat automatische OCR-herkenning geen 100% correcte uitvoer oplevert. Men zou daarom graag over indicatieve gegevens beschikken over de mate waarin de OCR voor bepaalde periodes of kranten is uitgevoerd. De bruikbaarheid van de krantendatabank voor onderzoek zal toenemen als de OCR-uitvoer verbeterd wordt. De onderzoekers zouden graag zien dat hieraan blijvend aandacht besteed wordt. Ook mogelijkheden om handmatig de uitvoer aan te passen worden in dit verband genoemd.

Onderzoekers vinden het van belang om hulpmiddelen te kunnen gebruiken waardoor ze meer zoekresultaten krijgen. Fuzzy zoeken en het opnemen van lijsten zoals namenlijsten, geografische lijsten en dergelijke worden hier genoemd.

Ook het koppelen van zoekmogelijkheden aan externe databanken, bijvoorbeeld aan Google Maps of genealogische databanken, wordt als wenselijk gezien.

Van belang voor onderzoekers is ook om snel de juiste informatie te kunnen vinden en deze in een bepaalde context te kunnen plaatsen: gaat het hier om een artikel op de voorpagina, een opiniërend artikel, een achtergrond artikel? De artikelen moeten het liefst afzonderlijk doorzoekbaar zijn.

De deelnemers willen ook snel inzicht krijgen in de waarde die het betreffende artikel heeft voor het onderzoek. Zo wordt geopperd om bijvoorbeeld automatisch gegenereerde samenvattingen te presenteren waarin door middel van een aantal trefwoorden de essentie van het artikel wordt weergegeven.

In dit kader wordt ook genoemd dat men gebruik wil maken van een geavanceerde zoekfunctionaliteit zoals *proximity search*. Deze functionaliteit moet op een eenvoudige manier beschikbaar zijn zodat onderzoekers uitgenodigd worden er gebruik van te maken. Een klein deel van de onderzoekers is bekend met allerlei vormen van geavanceerde zoekfunctionaliteit. Deze moet echter expliciet aangeboden worden zodat er beter gebruik kan worden gemaakt van de mogelijkheden die *full text search engines* bieden.

Technieken die voortkomen uit het onderzoek naar *text mining* bieden voor sommige onderzoekers goede mogelijkheden om betere zoekresultaten te vinden. Omdat het om een groot corpus aan tekstuele gegevens gaat zijn deze relevant om de juiste informatie te vinden. Deze technieken moeten zeker in een later stadium gebruikt kunnen worden. Het heeft echter niet direct de hoogste prioriteit. Die ligt in eerste instantie bij een zo correct en betrouwbaar mogelijke ocr-uitvoer.

5.3 Achtergrondinformatie

Onderzoekers hebben graag de beschikking over contextuele gegevens: het verspreidingsgebied van de krant, de oplage in een bepaalde periode. Wat voor soort lezers betrof het hier? Geloof, opleiding en dergelijke zijn dan van belang. Deze contextuele gegevens zijn van belang om de informatie in een kader te kunnen plaatsen. In dit kader werd ook geopperd om biografieën van journalisten direct te koppelen aan de krant.

5.4 Presentatie

Onderzoekers willen graag verschillende mogelijkheden voor de presentatie kunnen benutten. Artikelen in een bepaalde volgorde: geografisch, tijd, krant. Artikelen in hun context: hele pagina, hele krant. Het is wenselijk om zoekresultaten op te kunnen slaan, en daarnaast om gevonden resultaten in een *winkelwagentje* op te kunnen slaan zodat ze in een later stadium gedownload kunnen worden.

Men wil contextuele informatie bij de resultaten kunnen opvragen. Men wil zowel de pdf als de *full text* kunnen bekijken en downloaden. De gevonden resultaten kunnen ook getoond worden als een netwerk, zodat relaties tussen verschillende artikelen en tijdsreeksen goed zichtbaar worden. Een artikel is een reactie op een ander artikel of een vervolg.

5.5 Analyse

De wijze waarop onderzoekers de gevonden informatie analyseren verschilt per onderzoeker. Soms is het slechts nodig om het artikel *online* te kunnen lezen. Anderen willen een selectie van de artikelen kunnen downloaden. Voor sommigen volstaat een pdf als *downloadbaar* item. Anderen willen het artikel als xml bestand met alle relevante structurele gegevens.

Voor sommigen volstaat een kleine selectie van artikelen. Anderen zijn geïnteresseerd in grote hoeveelheden data. Sommige onderzoekers willen de informatie downloaden om zo hun eigen *analyse tools* te kunnen gebruiken. Anderen willen graag de mogelijkheid hebben om *analyse tools* op het corpus los te kunnen laten.

Sommige onderzoekers willen annotaties kunnen maken bij artikelen. Deze annotaties moeten dan op een gepersonaliseerde manier worden opgeslagen en te downloaden zijn. Anderen opperen de mogelijkheid om annotaties op een gedistribueerde manier te kunnen koppelen aan artikelen in de krantendatabank.

Innovatief onderzoek is mogelijk door gebruik te maken van alle mogelijkheden die ICT op dit moment en op de lange termijn biedt. Er is echter een grote kloof tussen de kleine groep onderzoekers die zich op dit moment met deze vorm van onderzoek bezighoudt en de meerderheid van onderzoekers die op meer traditionele wijze onderzoek doet. Enige educatie van deze gebruikers, het laten zien welke mogelijkheden er zijn, kan het gebruik van de digitale krantendatabank voor innovatief onderzoek stimuleren.

5.6 Refereren

Voor onderzoekers is het van belang dat naar artikelen gerefereerd kan worden op basis van een uniek identificatiemechanisme. Dit mechanisme kan ook gebruikt worden om lijstjes van resultaten te bewaren.

5.7 Personalisatie en samenwerking

Onderzoekers geven aan dat kranten vaak slechts een onderdeel van de bronnen zijn die men voor het onderzoek gebruikt. *Collaboratory* functionaliteit die direct gekoppeld is aan de krantendatabank is dus niet echt gewenst. Wel hecht men er waarde aan dat voor persoonlijk gebruik men zoekresultaten zou kunnen opslaan, annotaties zou kunnen aanbrengen en dergelijke. Personalisatie van de krantendatabank is dan noodzakelijk.

6 Conclusies en aanbevelingen

Op basis van de verkenning kunnen een aantal conclusies worden getrokken. Deze conclusies leiden tot aanbevelingen met betrekking tot de gewenste functionaliteit voor het faciliteren van wetenschappelijk onderzoek. De hieronder opgesomde aanbevelingen zijn niet geprioriteerd naar belangrijkheid. Onderzoekers hechten zelf het meeste belang aan het full text doorzoekbaar maken van de krantendatabank. Daarnaast is het voor onderzoekers van belang om door middel van een kranten-ontologie gericht te kunnen zoeken naar informatie.

- Het gebruik van geavanceerde ICT-methoden en technieken is in de geestes- en maatschappijwetenschappen nog geen gemeengoed, in tegenstelling tot de meeste levenswetenschappen. Aanbevolen wordt om alle functionaliteit uitnodigend aan te bieden. De krantendatabank moet eenvoudig te gebruiken zijn. De meerwaarde van geavanceerde functionaliteit moet waar mogelijk gedemonstreerd worden zodat onderzoekers er ook in hun eigen onderzoek gebruik van gaan maken.
- Voor het gros van de onderzoekers is het een belangrijke stap voorwaarts als het bronnenmateriaal via het web wordt aangeboden en *full text* doorzoekbaar is. Aanbevolen wordt om blijvend aandacht te besteden aan het verbeteren van de OCR-uitvoer. De mogelijkheid, op woordniveau te kunnen zoeken in de krantendatabank vormt het fundament voor een goed en betrouwbaar gebruik van de kranten als bron.
- Onder onderzoekers in de geestes- en maatschappijwetenschappen zijn er die nog niet veel gebruik maken van ICT-methoden en -technieken, maar ook die allerlei geavanceerde methoden en technieken gebruiken. Aanbevolen wordt om twee soorten *interfaces* aan te bieden. Eén waarop de *mainstream* onderzoeker eenvoudig zijn weg kan vinden en wordt uitgenodigd om gebruik te maken van alle functionaliteit. Daarnaast een tweede waarop allerlei geavanceerde methoden en technieken worden aangeboden. Op deze manier kunnen beide categorieën op maat bediend worden.
- Onderzoekers willen de gevonden informatie snel kunnen duiden. Het is aan te bevelen om gebruik te maken van een ontologie waarin de structuur van kranten beschreven kan worden.
- De OCR-uitvoer zal in eerste instantie geen goede zoekresultaten kunnen garanderen. Het is daarom aan te bevelen om geavanceerde technieken te gebruiken als *fuzzy search*, namenlijsten, thesauri en technieken uit met name *text mining*. Deze technieken kunnen leiden tot een sterk verbeterde toegankelijkheid van de krantendatabank. De mate waarin tekens herkend worden door OCR-software zal verschillen per periode en krant. Onderzoekers willen daarvan graag een indicatie hebben. Het is dan ook aan te bevelen om informatie over de betrouwbaarheid van de ocr-uitvoer op een transparante wijze aan te bieden.
- De krantendatabank bevat veel informatie. Het is daarom aan te bevelen om de gevonden informatie op een efficiënte wijze beheersbaar te maken. Te denken valt aan het sorteren, groeperen van lijsten per krant, periode, rubriek en een combinatie van deze categorieën. Een goede structuur (ontologie) is hier wederom onontbeerlijk. Individuele zoekresultaten moeten in verschillende granulariteit (artikel, pagina, krant) en in verschillende formaten (pdf, full text, image, xml) worden aangeboden. Het moet tevens mogelijk zijn om grote hoeveelheden materiaal in *bulk* te kunnen downloaden.
- Onderzoekers willen graag annotaties aan kunnen brengen. Het is daarom aan te bevelen om bepaalde functionaliteit van de krantendatabank gepersonaliseerd aan te bieden. Te denken valt hier aan het kunnen aanbrengen van annotaties, het bewaren van zoekopdrachten, het kunnen aanbrengen van relaties tussen artikelen en dergelijke.

- Onderzoekers willen op eenduidige wijze kunnen refereren aan informatie in de krantendatabank. Het is daarom aan te bevelen om bijvoorbeeld artikelen en foto's door middel van persistente *identifiers* te identificeren en lokaliseren.
- Voor longitudinaal en vergelijkend onderzoek moeten analyses mogelijk zijn op grote hoeveelheden data. Daarnaast moet het mogelijk zijn om geavanceerde tools te kunnen loslaten op de informatie in de krantendatabank. De krantendatabank zou zo ingericht moeten worden dat externe analysetools door middel van een eenduidige *application programming interface* (API) gebruikt kunnen worden. Daarnaast moet het mogelijk zijn om grote hoeveelheden geselecteerde informatie op een gestandaardiseerde wijze te kunnen downloaden.
- Achtergrondinformatie over de kranten of artikelen is voor onderzoekers van belang om de informatie contextueel te kunnen duiden. Het is aan te bevelen om externe bronnen te koppelen aan de krantendatabank. Te denken valt aan bronnen met informatie over verspreidingsgebied, aantal lezers en signatuur van een krant en biografische gegevens van journalisten.
- Onderzoekers zien op dit moment nog geen meerwaarde van een zogenaamde *collaboratory* functionaliteit. Het is aan te bevelen om wel rekening te houden met de mogelijkheid dat deze functionaliteit in een later stadium wenselijk wordt bevonden.
- Uit de verkenning komt naar voren dat kranten grofweg op drie verschillende manieren als bron gebruikt worden: voor het vergaren van achtergrondinformatie; voor longitudinaal/vergelijkend onderzoek en als tekstcorpus voor het onderzoek naar tekst en taal. Het is aan te bevelen om de functionaliteit van de krantendatabank af te stemmen op deze drie soorten van gebruik.
- Onderzoekers verwachten met name beter en efficiënter longitudinaal en vergelijkend onderzoek te kunnen doen met behulp van de krantendatabank. Wanneer het mogelijk is om relaties tussen artikelen aan te brengen en te visualiseren, kan dat de kwaliteit van het onderzoek flink ten goede komen. Daarnaast kan gedacht worden aan functionaliteit zoals woordthesauri die hedendaagse termen koppelen aan oude, of historisch-geografische functionaliteit die hedendaagse geografische aanduidingen koppelt zijn aan eerdere benamingen en afbakeningen in de tijd.

Bijlagen

Bijlage 1: Lijst van disciplines die refereren aan kranten in onderzoek

Bijlage 2: Aanvullende zoekfunctionaliteiten afkomstig uit interviews

Bijlage 3: Lijst van gevonden functionaliteiten in digitale kranten

Bijlage 4: Voorbeeld van een krantontologie

Bijlage 5: Relevante onderzoeksprojecten

Bijlage: Lijst van disciplines die in onderzoek refereren aan kranten

| Discipline | # referenties per discipline | # artikelen waarin ook andere kranten geciteerd worden |
|--|------------------------------|--|
| NRC (50) | | |
| Anthropology, geography, sociology | 2 | |
| Area studies, history, history & philosophy of science | 5 | |
| Business, ethics, planning & development, urban studies | 4 | 3 |
| Chemistry (analytical) | 1 | |
| Demography, ethnic studies | 1 | |
| Economics, geography, Environmental sciences, Environmental studies | 9 | 2 |
| Ethics, social issues, social sciences, biomedical, Medicine (general & internal), Medicine(legal), oncology, Public, environmental, & occupational health | 5 | 2 |
| Information science & library science | 1 | |
| International relations, political science, social issues | 3 | 1 |
| law | 1 | 1 |
| Management, social sciences | 2 | 1 |
| Marine & freshwater biology, microbiology | 2 | |
| Pharmacology & pharmacy, plant sciences | 1 | |
| Psychiatry (substance abuse), Psychology (social), Psychology, (mathematical), social sciences, mathematical methods | 3 | |
| Public administration, behavioral sciences (substance abuse), Social issues, Social sciences (mathematical methods), (interdisciplinary), sociology. | 8 | 4 |
| Woman's studies | 2 | 1 |
| Volkskrant (23) | | |
| Applied linguistics, sychology (social), psychiatry, | 3 | |
| Area studies | 1 | |
| Business, planning & development | 1 | |
| Communication | 2 | 1 |
| Economics, geography | 1 | |
| Education & educational | 1 | 1 |
| Energy & fuels, engineering, petroleum | 1 | |
| History & philosophy of science, agriculture, multidisciplinary, environmental sciences. | 2 | 1 |
| Language & linguistics, law | 1 | |
| Medicine (general & internal), public, environmental & occupational health, social sciences, biomedical, medicine (legal) | 4 | 2 |
| Metallurgy & metallurgical engineering | 1 | |
| Political science | 1 | 1 |
| Sociology, social issues | 2 | |

Eindrapport

| | | |
|--|---|---|
| Women's studies | 2 | |
| AD (3) | | |
| Biotechnology & applied microbiology, environmental sciences | 1 | |
| Political science | 1 | |
| Sociology, social sciences, interdisciplinary, social issues | 1 | 1 |

Bijlage 2: Aanvullende zoekfunctionaliteiten afkomstig uit interviews

Ideeën van geïnterviewden

Vinden

- Aantal abonnees in een bepaalde periode
- Woonplaats abonnees in een bepaalde periode
- Per editie in periode
- Gerelateerde artikelen
 - artikelen als reactie op
 - ingezonden brieven
 - artikelen die op andere pagina verder
- Automatisch toekennen onderwerpcategorie
- Hiërarchie van onderwerpen
- Ook buiten categorieën om zoeken
- Performance --> snelheid is belangrijk(st)
- Beperkingen aan kunnen geven
- Per krant
- Over meerdere kranten
- Alternatieve zoekterm op basis van eigen zoekterm bieden
- Met steekwoord ook foto's kunnen vinden
- Nieuws versus achtergrond artikelen
- Titelwoord versus artikelwoord

Selecteren

- Op basis van kwaliteit van de tekst (OCR)
 - onderzoeker kent schaalscore toe
 - foutenpercentage per jaargang
- Op basis van relevantie
 - tijd die onderzoeker op link doorbrengt

Analyseren

- Het aantal maal dat een bepaalde combinatie van woorden voorkomt
 - Werkwoord; zelfstandig naamwoord; plaats in de zin
- Niet gebonden aan functionaliteit van de Digitale Databank voor Dagbladen
 - alles kunnen downloaden, eigen tools
 - gesloten systeem is gemiste kans
 - uploaden en delen van eigen tools
 - netwerkanalyse, personen samen genoemd
- Krant als inventarisatie op onderwerp
- Altijd inzicht in analysetechniek houden
- Onderzoeker idee van totale controle
- (her)Categoriseren faciliteren
- Lezers review van artikel op basis van:
 - kwaliteit analyses door auteur
 - kwaliteit / hoeveelheid info

- in hoeverre overzichtsartikel?
- in hoeverre is artikel vormend idee voor onderzoeker
- Zien hoe vaak zoekterm wordt gebruikt

Bewerken

- Annotaties
- Highlighten
- Kunnen printen

Presenteren

- Trefferlijst zichtbaar naast artikel
- Artikel als thumbnail
- Tabbladen
 - essentie op eerste blad
 - extra's op bladen daar achter

Bewaren

- Persoonlijke bookmark collectie

Terugvinden

- Persoonlijke bookmark collectie
- Iedere pagina unieke url

7 Bijlage 3: Lijst van gevonden functionaliteiten in digitale kranten

Zoekopties

| | |
|--------------|--|
| Keywords | Er kan worden gezocht op voorkomens van keywords in tekst |
| Namen | Er kan specifiek worden gezocht op namen |
| Auteur | Er kan specifiek worden gezocht op auteur |
| Periode | Er kan worden gezocht binnen een bepaalde periode |
| Locatie | Er kan worden gezocht binnen een bepaald gebied |
| Sectie | Er kan worden gezocht binnen bepaalde delen van de krant |
| Type | Er kan worden gezocht binnen bepaalde onderdelen van de krant (titel, advertentie, nieuws, etc.) |
| Artikelsoort | Er kan worden gezocht binnen een bepaald soort artikelen (economie, cultuur, etc.) |
| Inhoud | Er kan worden gezocht binnen de inhoudsopgave |

Zoekmechanismen

| | |
|-----------|---|
| Boolean | De queries kunnen worden gecombineerd met AND, OR en NOT |
| Fuzzy | Er kan worden gezocht op resultaten die klinken als een keyword Er kan worden gezocht op resultaten die dezelfde betekenis hebben als een keyword Er kan worden gezocht op resultaten die een vervoeging zijn van een keyword |
| Exact | Er kan worden gezocht op exacte woorden (denk aan accenten etc.) |
| Wildcards | Er kunnen wildcards worden gebruikt zoals * of ? |
| Near | Er kan worden gezocht op woordcombinaties waarbij de afstand tussen de woorden kan worden gespecificeerd |
| Sorteren | De resultaten kunnen worden gesorteerd |

Presentatiemogelijkheden

| | |
|--------------|--|
| PDF | Het artikel of de gehele pagina wordt als PDF bestand getoond |
| Image | Het artikel of de gehele pagina wordt als TIFF bestand getoond |
| Tekst | Het artikel of de gehele pagina wordt als tekst (al dan niet met opmaak) getoond |
| Highlighting | De artikelen kunnen worden gehighlight wanneer de muis erover beweegt om zodoende dit artikel te kunnen selecteren |
| Flash | Het artikel of de gehele pagina wordt als Flash-animatie getoond |

Bijlage 4: Voorbeeld van een krantontologie

Een uitwerking van de opbouw van een krant:

Krant

Bestaat uit:

Naam
Editie
Katern *
Pagina *

Kenmerken:

Paginanummer
krantnaam
Voorkant / Achterkant (van een katern)

Item *

Kenmerken:

Positie / afmeting

Kan zijn:

Nieuws
Advertentie
Recensie
Column
Inhoudsopgave
Weer
TV Gids
Colofon
Strip
Puzzel

Bevat:

Plaats
Auteur
Abstract
Foto / Grafiek / Schema / ...
Bijschrift
Tekst
Quote
Toelichting / Uitwijding

* = 0 of meerdere

Bijlage 5: Relevante onderzoeksprojecten

Onderzoeksprojecten uit het CATCH-programma

Multiple-collection Searching Using Metadata

Programma: CATCH
 Naam: Museum
 Website: <http://staff.science.uva.nl/~kamps/museum/>
 Omschrijving: Onderzoek naar de integratie van meerdere (sub-)collecties welke verschillende talen, metdata, etc. bevatten. Betreft een theoretisch onderzoek naar de effectiviteit hiervan.
 Toepassing: Collectie-integratie

Reading Images in the Cultural Heritage

Programma: CATCH
 Naam: Ritch
 Website: <http://www.rich.unimaas.nl/>
 Omschrijving: Het gebruik van kunstmatige intelligentie om archeologische vondsten te herkennen en semi-automatisch te classificeren op basis van vorm, materiaal en decoratie.
 Toepassing: Beeld-analyse

Semantic Interoperability to Access Cultural Heritage

Programma: CATCH
 Naam: Stitch
 Website: <http://www.cs.vu.nl/STITCH/>
 Omschrijving: Ontwikkelen van theorie, methoden en tools om metadata-operabiliteit tussen de verschillende vocabularies te creëren m.b.v. semantische links (zoals het ontology-mapping-probleem van het semantische web).
 Toepassing: Collectie-integratie

Mining for Information in Texts from the Cultural Heritage

Programma: CATCH
 Naam: Mitch
 Website: <http://ilk.uvt.nl/mitch/>
 Omschrijving: Het ontwikkelen van technologie om een collectie minder-gestructureerde documenten en (manual) databases doorzoekbaar te maken door het automatisch linken van informatie in de verschillende bronnen.
 Toepassing: Collectie-integratie

Multimedia Analysis for Cultural Heritage

Programma: CATCH
 Naam: MUNCH
 Website: <http://ilps.science.uva.nl/munch/index.html>
 Omschrijving: Onderzoek naar automatische analyse van multimedia (plaatjes) archieven om bladeren en zoeken mogelijk te maken.
 Toepassing: Beeld-analyse

Cultural Heritage Information Personalization

Programma: CATCH
 Naam: CHIP
 Website: <http://www.chip-project.org/>

Omschrijving: Ontwikkeling van technieken om toegang tot cultureel erfgoed te personaliseren op basis van ‘aanbevelingen. Hiervoor zal het semantische web worden gebruikt.
Toepassing: Collectie-integratie dmv personalisering

Charting the Information Landscape Employing Context Information

Programma: CATCH
Naam: Choice
Website: <http://www.nwo.nl/catch/choice>
Omschrijving: Onderzoek naar het semi-automatisch semantisch annoteren van video, plaatjes en boeken met semantische categorieën uit gestandaardiseerde metadata-repositories zoals domein-thesauri en ontologieën
Toepassing: Beeld-analyse

Access to Oral History

Programma: CATCH
Naam: Choral
Website: <http://hmi.ewi.utwente.nl/choral>
Omschrijving: Het ontsluiten van gesproken historische collecties door middel van spraakherkenning en information retrieval.
Toepassing: Audio-analyse

Internationale Onderzoeksprojecten

Semantically Enhanced, Multifaceted Collaborative Access to Cultural Heritage

Programma: EU IST Programme
Naam: Mosaica
Website: <http://www.mosaica-project.eu/>
Omschrijving: Het ontwikkelen van een proof-of-concept dat gedistribueerde culturele collecties “knowledgebased” ontsluit en op een interactieve creatieve manier presenteert
Toepassing: Collectie-integratie

Query and context based visualization of time-spatial cultural dynamics

Programma: FP6
Naam: Qviz
Website: <http://qviz.eu/>
Omschrijving: Onderzoek en ontwikkeling van een framework voor het doorzoekbaar maken en presenteren van archieven op basis van een tijdschaal en andere schalen (bijv. locatie).
Toepassing: Toegang

Multilingual/Multimedia Access to Cultural Heritage

Programma: Europees
Naam: MultiMatch
Website: <http://www.multimatch.eu>
Omschrijving: Doel van het project is om cultureel erfgoed online doorzoekbaar en presenteerbaar te maken onafhankelijk van media-type of taal.
Toepassing: Collectie-integratie

Image-based Navigation in Multimedia Archives

Programma: FP6
Naam: Imagination
Website: <http://www.imagination-project.org/>

Omschrijving: Ontwikkeling van een systeem waar plaatjes en delen van plaatjes aan elkaar zijn gelinkt door middel van ontologieën waardoor hier doorheen gebrowsed kan worden
Toepassing: Collectie-integratie

Text Analysis Portal for Researchers

Programma: Canadees
Naam: TAPOR
Website: <http://www.tapor.ca>
Omschrijving: Het ontwikkelen van een portal voor linguïsten welke representatie, analyse en samenwerking met teksten mogelijk maakt. Op <http://portal.tapor.ca> worden op dit moment text-tools aangeboden.
Toepassing: Collaboration

Networked Infrastructure for Nineteenth-century Electronic Scholarship

Programma: Andrew W. Mellon Foundation
Naam: Nines
Website: <http://www.nines.org>
Omschrijving: Stelt teksten en plaatjes beschikbaar en maakt het mogelijk eigen collecties aan te leggen, annotaties te maken, etc.
Toepassing: Collaboration

General Architecture for Text Engineering

Programma:
Naam: Gate
Website: <http://www.gate.ac.uk>
Omschrijving: biedt 'Information Extraction', het automatisch vertalen van tekst naar onambigue informatie.
Toepassing: Tekst-analyse

TextGrid

Programma: Bundesministerium für Bildung und Forschung
Naam: Textgrid
Website: <http://www.textgrid.de>
Omschrijving: Binnen een grid-context worden tools ontwikkeld voor bewerking, annotatie, analyse en publicatie.
Toepassing: Collaboration

Discovery

Programma: eContentPlus
Naam: Discovery
Website: <http://www.discovery-project.eu>
Omschrijving: Het opzetten van een collaboratory / semantic web environment om een verzameling filosofische werken
Toepassing: Collaboration

Overig onderzoek

Nog enkele soorten onderzoek of techniek welke niet direct als project zijn aan te wijzen

Elaborate

Programma:
Naam: Elaborate / Womans Writers
Website: <http://www.e-laborate.nl>
Omschrijving: Veel gebruikte online tool voor het transcriberen, annoteren, categoriseren, etc. van tekstmateriaal
Toepassing: Categoriseren

Tekstanalyse

Programma: Algemeen
Naam: Tekstanalyse
Website:
Omschrijving: Een voorbeeld van tekstonderzoek is het herkennen / onderscheid maken van teksten van verschillende auteurs.
Toepassing: Tekst-analyse

Semantic Web

Programma: TNO
Naam: Semantic Web
Website:
Omschrijving: Bij TNO wordt aan een systeem gewerkt dat op basis van woordovereenkomst semantische kennis opbouwt.
Toepassing: Analyse, Vinden

Text Genetics / Stemmaties

Programma:
Naam: Text Genetics / Stemmaties
Website:
Omschrijving: Onderzoek naar hoe teksten zijn ontstaan en/of van elkaar afstammen
Toepassing: Analyse

Collate

Programma: Itsee
Naam: Collate
Website: <http://www.itsee.bham.ac.uk>
Omschrijving: Een collaberatory voor het samenwerken met textdata. In een GRID-omgeving.
Toepassing: –