

Peter Doorn

Modern monnikenwerk: digitalisering van historische bronnen

‘Invoeren van gegevens blijft een tijdrovende en vaak geestdodende bezigheid’, schreef George Welling in *Groniek* 149. Met behulp van scanners kun je wel snel digitale plaatjes van historische bronnen maken, maar dat biedt volgens hem weinig soelaas, want met gegevens in plaatjes kun je niet rekenen. Bij het Nederlands Historisch Data Archief (NHDA, dat in 1997 opging in het Nederlands Instituut voor Wetenschappelijke Informatiediensten - NIWI) is sinds de vroege jaren negentig ervaring opgedaan met het digitaliseren van historische bronnen. In dit artikel gaat Peter Doorn in op de stellingen van Welling en geeft hij een weergave van enkele ervaringen met scannen, optische tekenherkenning en handmatige invoer van bronnen, in het bijzonder van de Nederlandse Volkstellingen.

1. Inleiding

Bij het college voortgezette statistiek, dat ik vele jaren aan Leidse studenten geschiedenis heb gegeven, zat stevast een voorbeeld over het voordeel van steekproeven bij historisch onderzoek: stel dat we beschikken over een bron met gegevens over 100.000 eenheden. Er zijn vijf minuten nodig om de gegevens over één eenheid over te nemen en in te voeren. Het invoeren in de computer van het gehele bestand kost derhalve 500.000 minuten. Als je vijf dagen per week acht uur lang gegevens invoert, zonder vakantie te nemen of koffie te drinken, betekent dit dat je bijna vijf jaar bezig zou zijn om alle gegevens in te voeren. Als je een steekproef van 1.000 eenheden zou trekken, dan heb je aan twee weken data-invoer voldoende.

Hiermee wil ik niet zeggen dat het trekken van een steekproef de panacee is voor het historisch bronnenonderzoek, maar wel dat je van tevoren moet nadenken over de bedoeling van je onderzoek en hoe je dit het beste kunt aanpakken. Het feit dat zoveel historici een groot gedeelte van hun kostbare onderzoekstijd besteden aan betrekkelijk eenvoudig typewerk vormde voor het toenmalig Nederlands Historisch Data Archief (NHDA) aanleiding om in 1989 een eerste verkenning uit te voeren naar de mogelijke inzet van scanning en optische tekenherkenning (Engels: *Optical Character Recognition*, OCR) bij het invoeren van bronnen.¹

In de ruim tien jaar die sindsdien verstreken zijn, heeft het NHDA tal van projecten uitgevoerd om historische bronnen te digitaliseren, ontsluiten en archiveren. Welke technieken het beste konden worden gebruikt om de gegevens vanaf het papier in digitale vorm te krijgen is afhankelijk van enerzijds de bron en anderzijds van het oogmerk van het digitale bestand. Bijna alle projecten zijn in opdracht van of in samenwerking met andere partijen uitgevoerd. Deze opdrachtgevers en samenwerkingspartners omvatten vele nationale en zelfs internationale historische organisaties: Instituut voor Nederlandse Geschiedenis, Internationaal Instituut voor Sociale Geschiedenis, Nederlands Instituut voor

¹ L.J. Touwen, ‘Scanning van historische bronnen: valkuilen en mogelijkheden’, in: *Nederlands Historisch Data Archief I: eindverslag van een pilot project*, samengesteld door P.K. Doorn et al. (Amsterdam/Utrecht 1990), p. 23-36; René van Horik, *Van beeldpunt tot betekenis: scanning en optische tekenherkenning van gedrukt historisch bronnenmateriaal* (Amsterdam, 1992). *Optical Character Recognition in the historical discipline: proceedings of an international workshop* (Halbgraue Reihe zur historischen Fachinformatik) (Sankt Katharinen, 1993).

Oorlogsdocumentatie, Algemeen Rijksarchief, diverse stadsarchieven (waaronder dat van Antwerpen en de London Metropolitan Archives), de Koninklijke Bibliotheek, verscheidene Universiteitsbibliotheken en universitaire geschiedenisafdelingen, postdoctorale opleidingen en onderzoeksscholen. Ook met het Centraal Bureau voor de Statistiek werd een omvangrijk digitaliseringsproject uitgevoerd: de digitalisering van de Nederlandse Volkstellingen.

In een vorig nummer van *Groniek* bekritiseerde George Welling dit project in niet geringe mate². Mijns inziens was zijn kritiek dermate onjuist en onterecht, dat ik hem hier uitvoerig van repliek dien. Bovendien zal ik juist het Volkstellingenproject gebruiken als voorbeeld, omdat hierin tal van ervaringen zijn opgedaan, die ook voor andere digitaliseringsprojecten relevant zijn.

2. Welling's kritiek en mijn repliek

Welling vindt de digitale editie van de Volkstellingen van CBS en NIWI een 'treurig' product, omdat deze louter afbeeldingen van eerder gedrukte pagina's zou bieden. Een eenvoudig zoekprogramma helpt wel om snel een bepaalde pagina op te zoeken, maar je zou de gegevens eerst zelf moeten invoeren alvorens je ze kunt analyseren. Bovendien vindt hij de kwaliteit van de digitale beelden (*images*) niet altijd voldoende om OCR toe te passen. En omdat OCR toch niet mogelijk is, had een compacter opslagformaat gekozen kunnen worden, zodat alle informatie op één CD-ROM zou passen in plaats van op vijf.

Het is Welling blijkbaar ontgaan dat het volkstellingenproject, behalve de bewuste set van vijf CD-ROM's met images, nog diverse andere producten heeft opgeleverd. Van meer dan 10.000 pagina's met volkstellingspublicaties (ongeveer een kwart van het totaal aantal gepubliceerde pagina's sinds 1795) zijn direct analyseerbare databases en vrij doorzoekbare tekstbestanden gemaakt, zowel op het Web als op CD-ROM. De gemaakte keuzen en selecties worden nader toegelicht in paragraaf 4 en verder.

De kwaliteit van de images is ruim voldoende om OCR toe te passen. In het project is zelfs uitvoerig onderzoek gedaan naar de mogelijkheden om de tabellen optisch te herkennen. Niet de kwaliteit van de images vormt het probleem, maar de staat van de bron en de structuur van de gegevens. Voor bestaande OCR-programmatuur blijkt het zeer moeilijk om getallen op de juiste posities in rijen en kolommen te herkennen. De kwaliteit van de oorspronkelijke bron vormt voor sommige jaren ook een probleem. Een verantwoording van de werkwijze wordt geboden in paragraaf 6. Het gekozen opslagformaat biedt een efficiënte compressie zonder enig kwaliteitsverlies. Wel staan er ca. 200 kleurenimages op de CD's van uitklapkaarten, grafieken en dergelijke, die van te slechte kwaliteit zijn om optisch herkend te worden, maar dat zou voor kleurenplaten toch niet erg nuttig zijn. Het gaat hier om afgeleide images van digitale *masters*, die individueel vele megabytes beslaan. De digitale originelen zijn op speciaal verzoek beschikbaar bij het NIWI.

3. Digitalisering en data-archivering

De digitalisering van historische bronnen vormt eigenlijk slechts één stap uit een hele keten van activiteiten: om gegevens toegankelijk en begrijpelijk te maken is het noodzakelijk om informatie over de data toe te voegen (metadata en/of beschrijvende data-documentatie). De

² George M. Welling, 'Digitale bronnentranscripties: wie gaat ze invoeren?', in: *Groniek* 33-149 (2000) pp. 493-506.

structuur van de gegevens en hoe deze zich verhouden tot de oorspronkelijke bron moge aan de bewerker duidelijk zijn, voor eventuele raadplegers van de gegevens is dat doorgaans niet het geval. Bovendien is het zo dat eenmaal elektronisch toegankelijk gemaakte gegevens niet vanzelf eeuwig beschikbaar blijven. Er moet zorg gedragen worden voor de ‘digitale duurzaamheid’ van de gegevens, dat wil zeggen, voor de toegankelijkheid ervan op de lange termijn. Speciaal met dit doel is het Nederlands Historisch Data Archief opgericht.

Elders in zijn artikel laat Welling zich nogal laatdunkend uit over ‘toevallige bijproducten van historisch onderzoek, zoals die bijvoorbeeld bij het NHDA/NIWI gedeponereerd zijn’. Als voorbeeld noemt hij de pondtol registers van Elbing, een bestand dat door Thomas Lindblad bij het NHDA is gedeponereerd. Welling vindt de waarde van dergelijke bestanden betrekkelijk, omdat bij het aanleggen van bestanden ten behoeve van historisch onderzoek selecties en keuzes worden gemaakt, die de bruikbaarheid van de gegevens voor andere onderzoekers beperken.

Opmerkelijk is dat juist de door Welling genoemde dataset de basis vormt van één van de Rijks Geschiedkundige Publicatiën van het Instituut voor Nederlandse Geschiedenis.³ Het bestand heeft enerzijds gediend ter analyse van de Nederlandse scheepvaart en handel op het Oostzeegebied, anderzijds vormde het de basis voor een bronuitgave. Hoewel bij elektronische bronnenpublicaties meer gestreefd wordt naar het zo getrouw mogelijk digitaliseren van de oorspronkelijke bronnen dan bij onderzoeksbestanden, is het onderscheid niet zo scherp als Welling doet voorkomen. Niet alleen bij het aanleggen van bestanden ten behoeve van onderzoek, ook bij digitale bronnenpublicaties moeten keuzes worden gemaakt. We zullen zo direct zien dat ook Welling zelf niet aan dergelijke keuzes ontkomt.

Het NIWI is op dit moment betrokken bij een groot digitaliseringsproject, waarbij eveneens afwegingen moeten worden gemaakt tussen (handmatige) data-entry aan de ene kant en scanning en ontsluiting van oorspronkelijke documenten in de vorm van images aan de andere kant. Hoewel er enkele miljoenen gulden en vele mensjaren arbeid met het project gemoeid zullen zijn, is het bij voorbaat zonneklaar dat het onmogelijk is om alle beschikbare bronnen volledig te digitaliseren. Overeenkomstig de wensen van de opdrachtgever en in overleg met de samenwerkingspartner is vooralsnog besloten om de nadruk te leggen op het creëren van databases. Het is echter onontkoombaar dat er selecties worden gemaakt van de in te voeren gegevens. Scans van de oorspronkelijke documenten zouden het voordeel hebben dat de volledige rijkdom van de bron kan worden getoond en gecontroleerd door de onderzoeker. Maar afbeeldingen van documenten zijn niet direct analyseerbaar; gestructureerde databases zijn dat wel, zoals ook Welling opmerkt.

In de rest van zijn artikel gaat Welling in op zijn eigen ‘toevallige bijproduct van historisch onderzoek’, de data-entry van de zogenaamde *Paalgeldregisters* (Havenboeken van de heffing van het Paalgeld). Alle inventieve invoerhulpjes ten spijt, slaagt Welling er niet in om de gegevens van alle 65 jaren waarvoor registers beschikbaar zijn, in te voeren. Na vele jaren van bloed, zweet en tranen lukt het hem om de gegevens voor 18 jaren geheel in te voeren en voor 29 jaren gedeeltelijk. Hoewel hij niet vermeldt hoe lang hij met dit monnikenwerk bezig is geweest, is mijn schatting toch dat het hem ettelijke arbeidsjaren moet hebben gekost.⁴ Opmerkelijk is, dat hij anderen verwijt gegevens niet compleet in te voeren, terwijl zijn eigen datasets ook lacunes vertonen. Ik wil die lacunes zelf in het geheel niet

³ J.Th. Lindblad, *Dutch entries in the pound-toll registers of Elbing 1585-1700*, Rijks Geschiedkundige Publicatiën, Grote Serie 225, 's- Gravenhage 1995.

⁴ Welling, ‘Digitale bronnen-transcripties’, p. 499 en verder. De verwerking van de gegevens van één jaar kostte al bijna een jaar aan werk. In doorlooptijd is Welling meer dan 10 jaar met zijn project bezig geweest.

bekritisieren. Integendeel, men kan zich afvragen of het voor het onderzoek van Welling noodzakelijk was om zoveel invoerwerk te doen.⁵ Zijn beschrijving van de gevolgde werkwijze werkt sterk de indruk dat er niet bepaald een plan ten grondslag lag aan zijn noeste arbeid. Overigens heb ik de indruk dat hij niet de enige historicus is met een zekere afkeer van een planmatige aanpak van het onderzoek. Een duidelijke strategie van dataverzameling, bijvoorbeeld het trekken van een steekproef, was wellicht effectiever geweest dan het volledig invoeren van de jaren 1742 en 1771-1787 en van de transatlantische handelsgegevens over de jaren 1788-1817.⁶ Het had Welling jaren van geestdodende arbeid kunnen schelen, die hij had kunnen besteden aan echt historisch onderzoek.

4. Digitalisering Nederlandse Volkstellingen

Het Centraal Bureau voor de Statistiek (CBS) en het NIWI werken sinds 1997 samen aan het project 'Digitalisering Nederlandse Volkstellingen 1795-1971'.⁷ De eerste algemene volkstelling in Nederland vond plaats in 1795 ten tijde van de Bataafse Republiek. Sedert 1829/1830 is er sprake van periodieke tellingen, die eens in de 10 jaar werden gehouden. De telling van 1940 werd in verband met WOII uitgesteld tot 1947. Na de veertiende telling in 1971 is in Nederland geen Volkstelling meer gehouden als gevolg van de toegenomen privacy-bewustheid (weigering tot medewerking) van de bevolking.

In totaal zijn in de laatste twee eeuwen een kleine 200 banden met ca. 42.500 bladzijden aan tabellen (en toelichtingen daarop) gepubliceerd. In de bibliotheek en het archief van het CBS bevinden zich daarnaast nog enkele honderdduizenden bladen met ongepubliceerd materiaal over de laatste drie Volkstellingen (1947, 1960 en 1971). Van de tellingen van 1960 en 1971 zijn ook digitale bestanden bewaard gebleven.

Nationale volkstellingen behoren tot de meest elementaire informatiebronnen over de toestand in een land. De volkstellingen bevatten een schat aan historische, sociaal-economische, demografische en culturele gegevens.⁸ Naast de omvang van de bevolking

⁵ G.M. Welling, *The prize of neutrality: trade relations between Amsterdam and North America 1771-1817 - a study in computational history* (Hilversum 1998).

⁶ De gegevens zijn op Internet beschikbaar, maar er moet worden opgemerkt dat de bij de bestanden beschikbare documentatie zo summier is, dat de gegevens voor een niet-ingewijde in de oorspronkelijke bronnen moeilijk te interpreteren zijn.

⁷ De digitalisering van de Nederlandse Volkstellingen is gesubsidieerd door het fonds Innovatie Wetenschappelijke Informatievoorziening (IWI) van SURF en door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Gebiedsbestuur voor de Maatschappij- en gedragswetenschappen en Wetenschappelijk Statistisch Agentschap).

⁸ De Nederlandse Volkstellingen 1795-1971 zijn digitaal beschikbaar op cd-rom en voor een deel op het Internet.

De cd-rom uitgave omvat twee sets cd-rom's:

Set 1: Data en publicatie volkstelling 1899. Twee cd-rom's. ISBN 90.6861.176.3, fl. 129,00. Deze uitgave bevat de volledige tekst van de Inleiding op de Volkstelling van 1899, alle gepubliceerde tabellen als StatLine-databases en digitale afbeeldingen van alle ca. 10.000 pagina's van de oorspronkelijke gedrukte uitgave. Met het bijgeleverde Statline-programma kan de gebruiker tabellen samenstellen voor een zelf gekozen selectie van gegevens.

Set 2: Publicaties volkstellingen 1795-1971. Vijf cd-rom's, ISBN 90.6861.177.1, fl. 249,00. Deze uitgave bevat digitale afbeeldingen van de ca. 42.500 pagina's van alle gedrukte publicaties van de Volkstellingen tussen 1795 en 1971.

De cd-rom's zijn te koop bij de boekhandel of rechtstreeks te bestellen bij:

Uitgeverij Stichting beheer IISG, Cruquiusweg 31, 1019 AT Amsterdam, Tel. (020) 668 58 66, fax (020) 665 64 11

De homepage van de Internet-publicatie is www.volkstelling.nl. Hier bevindt zich de volledige tekst

bevat de volkstelling doorgaans informatie over de structurele kenmerken van de bevolking van het land, zoals leeftijd, geslacht, burgerlijke staat, levensbeschouwing, huishoudenssituatie, beroepswerkzaamheid en nationaliteit. In diverse jaren is de volkstelling gecombineerd gehouden met een beroepstelling en een woningtelling.

De gepubliceerde tellingen bieden vaak zeer gedetailleerde informatie. De volkstelling 1899 (inclusief beroepstelling en woningtelling) omvat ca. 10.000 pagina's met tabellen. Eén tabelpagina bevat gemiddeld bijna 700 cellen. De gepubliceerde volkstelling 1899 telt in totaal dus zo'n zeven miljoen gegevens. Bijvoorbeeld, van iedere gemeente worden per buurt of wijk en type woonruimte gegevens gepresenteerd, die vaak tot op het individu herleidbaar zijn.

Van de Nederlandse Volkstellingen 1795-1971 is slechts een beperkt aantal exemplaren in Nederland aanwezig. Naast de bibliotheek van het CBS, beschikken ook de meeste universiteitsbibliotheken over een min of meer complete collectie. Veel van de gepubliceerde Volkstellingsboeken verkeren echter inmiddels in slechte staat.

Om uiteenlopende redenen is ervoor gekozen om de tellingen van 1899 als eerste ook inhoudelijk te digitaliseren. Dit wil zeggen dat niet alleen digitale opnamen (*images*) zijn gemaakt van alle bladzijden, maar dat alle gegevens zijn opgenomen in databases, die volledig doorzoekbaar en raadpleegbaar zijn. De volkstelling van 1899 is één van de meest uitvoerige tellingen. Zij staat bekend als een kwalitatief zeer goede telling, die was gecombineerd met een beroepstelling. Deze bevatte een zeer uitvoerige beroepenclassificatie, die het uitgangspunt heeft gevormd voor indelingen in de 20e eeuw. Er doen zich wel aanzienlijke verschillen voor met de beroepenclassificatie van 1889, die in historisch onderzoek veelal wordt gebruikt voor 19e eeuwse en eerdere sociaal-economische stratificaties.

Ten tweede ligt 1899 vrijwel halverwege de periode 1795-heden. De telling is enerzijds 'modern' van opzet, maar anderzijds vergelijkbaar met de eerdere tellingen in de 19e eeuw. Ten derde is het CBS opgericht in 1899. Het was de eerste telling die onder verantwoordelijkheid van het toen kersverse bureau werd uitgevoerd. Het eeuwfeest van het CBS in 1999 vormde een uitstekende gelegenheid voor een digitale heruitgave van de eerste CBS-telling.

5. Materiaalselectie

Een selectie van de te verwerken volkstellingsdelen 1795-1971 is gemaakt aan de hand van een vergelijking van de staat van conservering van de boeken in vijf bibliotheken. Behalve de delen met tabellen zijn ook belangrijke publicaties met informatie over de volkstellingen (zoals monografieën en inleidingen) verwerkt. In het algemeen bleek de drukkwaliteit van de geselecteerde boeken bevredigend. Wel zijn er delen die in alle bibliotheken dezelfde problemen vertonen (doordrukken achterzijde pagina's, krappe binding).

Op basis van een technisch vooronderzoek is vastgesteld dat een combinatie van microverfilming met microfilm scanning de meest efficiënte wijze van verwerking zou zijn. Microverfilming is bij meer digitaliseringsprojecten een zinvolle tussenstap. Niet alleen wordt een extra *hardcopy backup* gemaakt, ook het werkproces verloopt efficiënter dan bij scanning

van de Inleiding tot de Volkstelling van 1899 met bijbehorende tabellen en grafieken. Alleen de Webversie van deze Inleiding is volledig doorzoekbaar.

De StatLine-publicatie van de tabellen van de Volkstelling 1899 is tevens bereikbaar via de home page van het CBS www.cbs.nl.

van papier. Bij verwerking van zwart-wit drukwerk treedt geen noemenswaardig kwaliteitsverlies op. De microverfilming dient wel uiterst zorgvuldig en goed afgestemd op de scanning te geschieden. De verkleiningsfactor moet zo klein mogelijk zijn en de film moet contrastrijk zijn. Voor automatische scanning moeten de rolfilms zo lang mogelijk zijn en mogen deze geen lassen bevatten. Omdat de meeste tabellen doorlopen over twee pagina's, is er voor gekozen per *frame* twee pagina's op te nemen. De opnamen werden aangeleverd op 40 rolfilms met gemiddeld ruim 500 images per film. Afwijkende formaten (uitklaptabellen, kaarten en dergelijke) werden op een aparte film gezet.

De eerste volkstelling waarbij computerverwerking werd toegepast, was die van 1960. De bestanden zijn door het CBS enkele jaren geleden gedeponerd bij het Steinmetzarchief (thans is dit sociaal-wetenschappelijke data-archief onderdeel van het NIWI), nadat de ponskaarten c.q. oorspronkelijke tapes opnieuw waren ingelezen op het rekencentrum SARA van de Universiteit van Amsterdam. Bij de conversie en documentatie door het Steinmetzarchief is gebleken dat de bestanden lacunes vertonen en niet geheel overeenstemmen met de gepubliceerde volkstellingsresultaten. Het onderzoek naar de mogelijkheden van een 'digitale restauratie' van het bestand van de volkstelling 1960 is nog gaande.

De gegevens met bijbehorende documentatie van de volkstelling 1971 zijn bij het CBS nog aanwezig in het eigen computerarchief. Ook hier wordt nagegaan of en in welke vorm de bestanden van 1971 ter beschikking kunnen worden gesteld voor onderzoek. Hierbij moet de privacy van de ondervraagden gegarandeerd zijn, daarom zijn de gegevens reeds ten tijde van de telling geanonimiseerd.

6. Overtypen of automatisch herkennen?

Het NIWI is gespecialiseerd in het automatisch herkennen van historische documenten, maar bij tabellen doen zich extra problemen voor ten opzichte van lopende tekst. Bij tabellen is het essentieel dat de gegevens in de juiste rij en kolom terecht komen. De herkenning van de tabelstructuur vormt de grootste moeilijkheid bij het toepassen van OCR. Tabeltitels, opschriften, rij- en kolombeschrijvingen moeten worden onderscheiden van de inhoud van de tabel. De informatie in de tabellen is vaak hiërarchisch geordend. In de volkstellingen komen in de rijen geregeld vier of vijf hiërarchische niveaus voor (bijv.: gemeente > kom > buurt > woningtype). Ook de kolommen zijn dikwijls hiërarchisch gestructureerd. Ook doen zich nog complicaties voor, zoals voetnoten in tabellen, cijfers die bij het inbinden in de rug van het boek zijn verdwenen en gegevens die over verschillende cellen zijn samengenomen.

Het NIWI heeft onderzoek gedaan naar de mogelijkheden van het automatisch herkennen van tabellen.⁹ Het onderzoek heeft zich vooral gericht op het correct herkennen van de documentstructuur op basis van de uitvoer van OCR-software, die de coördinaten van ieder herkend teken levert. Bij de structurering wordt onderscheid gemaakt tussen rij- en kolombeschrijvingen en de eigenlijke inhoud van de tabel. Er is experimentele software ontwikkeld die, met gebruikmaking van een 'tweedimensionale grammatica' de tabelinformatie op een zodanige manier ontleedt, dat deze automatisch in de juiste rij en kolom wordt geplaatst. Deze software is overigens nog niet voor grootschalige productie toepasbaar, al is de aanpak veelbelovend.

Mediumconversie door middel van imaging

⁹ Mark Schraivesande, *The automatic recognition of tables* (Amsterdam, 1998). Internal report.

De microfilms van de volkstellingen zijn gescand met een Sunrise microfilmscanner.¹⁰ Alle images zijn gecontroleerd op kwaliteit en op ontbrekende pagina's. Op basis van de kwaliteitscontroles zijn diverse scans extra of opnieuw uitgevoerd. In het volkstellingsproject hebben de correcties en aanvullende scans (enkele honderden opnamen) een veelvoud van de benodigde tijd van het automatisch scannen van het volledige filmmateriaal gekost! Uitklapkaarten en pagina's met afbeeldingen die in kleur waren gedrukt zijn met een digitale Kontron-camera (in kleur) gescand.

Voor het opzoeken en raadplegen van de images van de tellingen van 1795-1971 zijn indexbestanden gemaakt. Hierin is onder andere het onderwerp op basis van de titelinformatie van de tabellen, teksthoofdstukken en andere belangrijke boekonderdelen opgenomen. De deel- en paginanummers van de tabellen zijn handmatig aan de images gekoppeld. Voor het opzoeken en raadplegen van de images van de tellingen van 1795-1971 is een eenvoudig zoekstelsel gemaakt. Er kan gezocht worden op het jaar van de telling, het soort of deel van de telling en op onderwerp. Het is ook mogelijk om door de images te bladeren.

Inhoudsconversie van de telling van 1899

De tekst uit de inleiding van de Volkstelling 1899 is bij het NIWI via scanning en OCR geconverteerd. Ook de kleine tabellen uit de inleiding en verschillende beroepenlijsten zijn automatisch herkend met OCR. Voor het overige zijn alle tabellen van de Volkstelling 1899 handmatig ingevoerd.¹¹ Er heeft controletoetsing plaats gevonden om het aantal invoerfouten zoveel mogelijk te beperken.

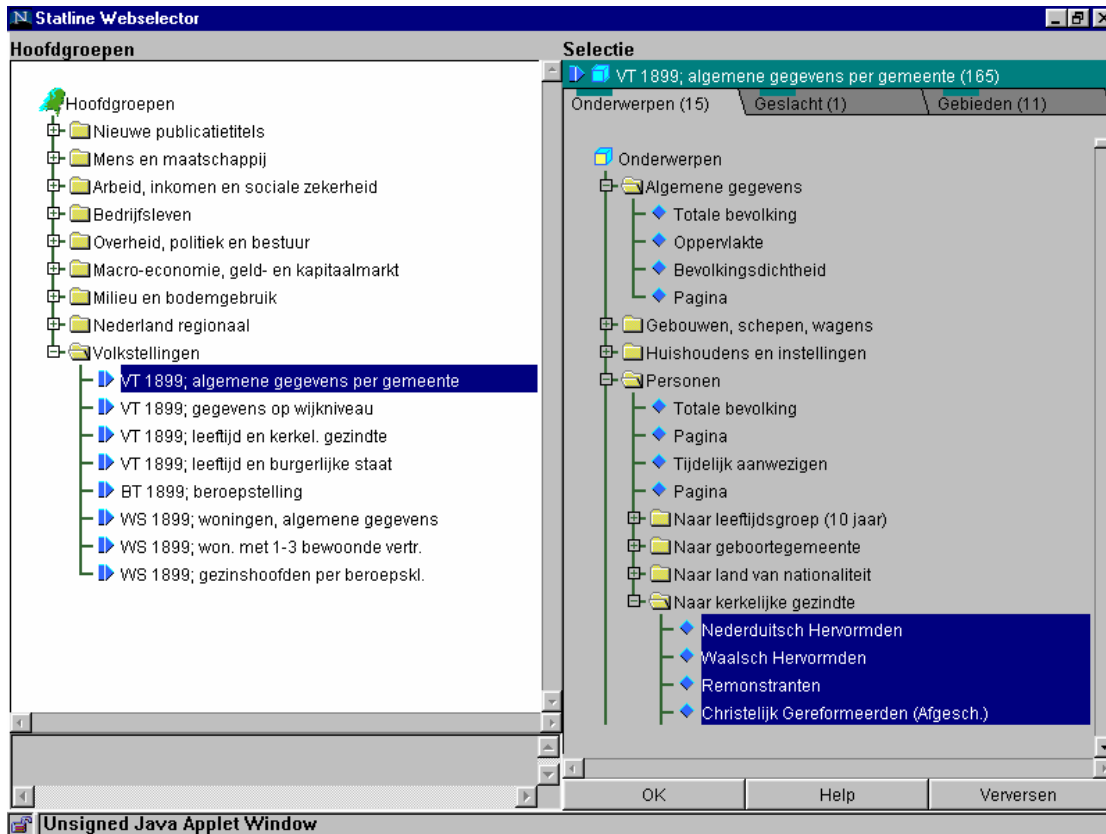
Bij de inhoudsconversie is er naar gestreefd om geen informatie verloren te laten gaan en om de gegevens op een zo 'brongetrouw' mogelijke manier over te nemen uit de publicaties. In principe is alle informatie uit de tabellen en toelichtende teksten gedigitaliseerd. De gebruiker heeft bovendien de digitale images als controlemiddel. Na voltooiing van de data-invoer van de tabellen van de volkstelling 1899 zijn controles op de juistheid van de gegevens in de database uitgevoerd. Belangrijkste instrument hierbij vormde het vergelijken van in de bron gegeven totalen met berekende totalen. Data-entry fouten zijn gecorrigeerd, maar bronfouten (druk- of optelfouten) niet.

De databestanden van de volkstelling 1899 zijn ontsloten met StatLine. Met dit stelsel maakt het CBS ook andere statistieken toegankelijk. StatLine is zowel beschikbaar op CD-ROM als op het World Wide Web.¹² Dit programma maakt het mogelijk dat de gebruiker zelf zijn tabellen samenstelt uit de beschikbare rij- en kolomvariabelen. De paginanummers van de oorspronkelijke tabel in de volkstellingsboeken zijn in de tabellen weer te geven, zodat de gebruiker de gegevens kan controleren met behulp van de images op CD-ROM.

¹⁰ De images zijn in zwart-wit gescand (geen grijswaarden) met een resolutie van 300 dpi (*dots per inch*) en opgeslagen als TIFF groep 4 (*lossless compression*). Op een deel van de images is beeldverbetering toegepast (*cropping, noise removal en deskewing*).

¹¹ Dat is gedeeltelijk gedaan bij het CBS (vestiging Heerlen) en gedeeltelijk, in opdracht van het CBS, bij IVA Data Entry Services BV te Rijswijk.

¹² www.cbs.nl



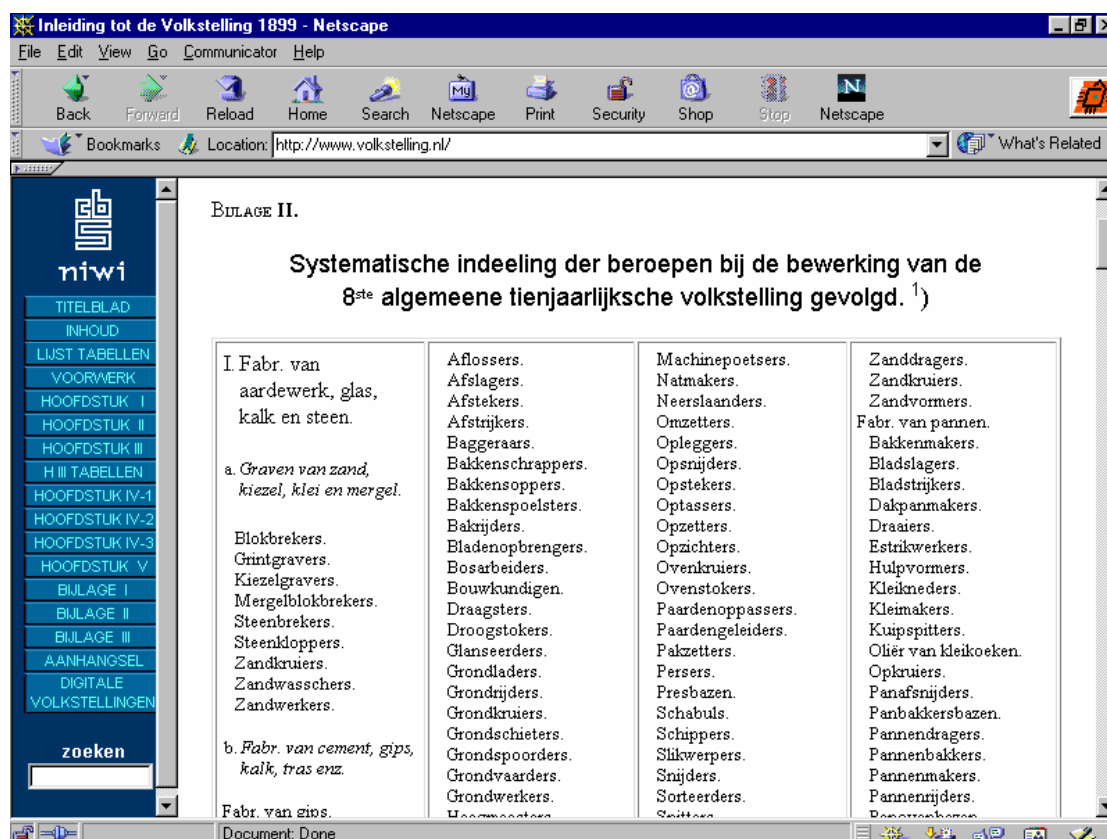
Figuur 1. Selectie van variabelen in de webversie van CBS-StatLine (www.cbs.nl)

		Personen							
		Naar kerkelijke gezindte							
		Nederduitsch Hervormden	Waalisch Hervormden	Remonstranten	Christelijk Gereformeerden (Afgesch.)	Doopsgezinden	Evangelisch Lutherschen	Hersteld Lutherschen	Behoorende tot de Gereformeerde Kerken
		<i>absoluut</i>							
Totaal	Groningen (prv.)	199 070	70	1 715	5 458	4 550	2 112	358	45 95
	Friesland	204 902	30	241	7 766	14 943	786	371	59 51
	Drenthe	112 193	23	1 001	2 461	641	241	67	18 16
	Overijssel	196 497	132	273	5 553	3 460	1 321	759	26 27
	Gelderland	314 414	576	1 796	3 508	1 701	3 105	950	27 31
	Utrecht (prv.)	134 647	607	1 181	2 339	1 088	2 857	599	18 96
	Noordholland	444 334	4 168	4 917	6 384	26 508	41 690	15 268	43 77
	Zuidholland	683 664	3 850	9 355	14 372	3 883	16 316	3 629	88 92
	Zeeland	127 147	117	66	5 805	645	875	167	21 28
	Noordbrabant	50 626	215	234	921	275	548	221	10 94
	Limburg	3 459	67	28	62	92	391	62	1

Figuur 2. Tabel samengesteld uit de Volkstelling van 1899 met CBS StatLine

Door de gebruiker gemaakte data-selecties kunnen eenvoudig worden opgeslagen om verder te worden bewerkt in een spreadsheet of statistisch pakket. De gegevens kunnen ook als grafiek of kaart (op gemeenteniveau) worden weergegeven.

De Inleiding 1899 is ontsloten via een speciale website. Deze is toegankelijk op CD-ROM en via het World Wide Web.¹³ De tekst en tabellen zijn geheel doorzoekbaar. De tabellen kunnen als afzonderlijke bestanden worden opgeslagen op een eigen schijf en vervolgens verder worden verwerkt in een spreadsheet-programma.



Figuur 3. Een deel van de beroepenclassificatie uit de gedigitaliseerde inleiding op de telling van 1899 (www.volkstelling.nl)

Geprobeerd is om de elektronische editie van de tekst van de Inleiding 1899 ook uiterlijk zo veel mogelijk te laten lijken op de oorspronkelijk gedrukte pagina's. Bij de lijsten en tabellen stond het gebruiksgemak bij verdere verwerking centraal (bijvoorbeeld analyse in een spreadsheet-programma). In de inleiding op de volkstelling 1899 is de variëteit aan tabellen zeer groot en blijkt ook de diversiteit aan fouten relatief groot te zijn.¹⁴ Dit hangt ongetwijfeld samen met het feit dat in deze relatief kleine, samenvattende tabellen, zeer uiteenlopende berekeningen en bewerkingen zijn uitgevoerd

De beroepenclassificatie van 1899

Voor de classificatie van de beroepen in 1899 is aanvankelijk uitgegaan van een beroepenoverzicht zoals geboden in een bijlage uit de Inleiding. Verondersteld werd dat deze

¹³ www.volkstelling.nl

¹⁴ Het gaat hier om fouten die door de samenstellers van de telling van 1899 gemaakt zijn, niet om data-entry fouten.

classificatie alle beroepen uit de telling zou omvatten. In de classificatie worden vier hiërarchische niveaus onderscheiden. Bij controles bleken echter aanzienlijke verschillen in beroepsomschrijvingen te bestaan tussen de classificatie en de formuleringen in de twaalf delen van de Beroepstelling. Hierop zijn ook de beroepsomschrijvingen uit de Beroepstelling van het Rijk als geheel gedigitaliseerd via scanning en OCR, alsmede de omschrijvingen uit een andere bijlage van de Inleiding. Na analyse van de verschillende lijsten is die uit het Rijksdeel van de beroepstelling genomen als groslijst voor de koppeling van de beroepenclassificatie met de cijfers in de tabellen. Toch werden tijdens de invoer nog steeds afwijkingen gevonden. Iedere afwijking van een beroepstitel die niet duidelijk een drukfout betrof, werd in het bestand geregistreerd. Op het totaal van ca. 100.000 records van de beroepstelling 1899 bleken ruim 2.500 beroepstitels niet in de groslijst voor te komen. Deze varianten zijn achteraf afzonderlijk behandeld en alsnog geklasseerd.

7. Conclusie

De eerste fase van de digitalisering van de volkstellingen is nu reeds ruim een jaar geleden voltooid. Zowel de CD-ROM's als de website worden intensief gebruikt bij historisch onderzoek. Binnenkort verschijnt een bundel artikelen met analyses van de volkstelling van 1899.¹⁵ Ook wordt er in diverse projecten verder gewerkt aan het invoeren en ter beschikking stellen van databases van de volkstellingen.¹⁶ De resultaten van het project rechtvaardigen een vervolg, waarbij zo mogelijk alle volkstellingen als databases ter beschikking komen, inclusief de naoorlogse handgeschreven tabellen in het archief en de bibliotheek van het CBS.

Bij een vervolgproject zal de onderlinge vergelijkbaarheid van de tellingen één van de grote problemen vormen. Ook koppelbaarheid van de volkstellingen aan andere grote onderzoeksbestanden op basis van individuele bevolkingsregistraties, zoals die van de Historische Steekproef Nederlandse bevolking (HSN) vormt een uitdaging. Het onlangs afgeronde project laat zien, dat juist redigeerwerk (correctie, beroepenclassificatie) zeer moeilijk en arbeidsintensief is. Gemeentelijke herindelingen, veranderende definities en uiteenlopende classificaties zullen een volledig consistente set digitale volkstellingen tot een illusie maken. Het zal een schone taak voor het historisch onderzoek zijn om na te gaan in hoeverre de volkstellingen in de loop der tijd vergelijkbaar gemaakt kunnen worden. En dan te bedenken dat de volkstellingen alleen maar het begin vormen van de onafzienbare hoeveelheden historische statistieken die in Nederland bewaard zijn gebleven.

We zien de laatste jaren een duidelijke schaalvergroting optreden bij digitaliseringsprojecten ten behoeve van de geesteswetenschappen. Bibliotheken, archieven en onderzoekers entameren steeds vaker omvangrijke projecten. Nog te vaak wordt echter uitgegaan van de collecties en te weinig van het onderzoeksbelang. Hoewel in diverse beleidsstukken met betrekking tot digitalisering van organisaties als NWO, de KNAW en SURF/IWI de nadruk wordt gelegd op digitalisering ten behoeve van het onderzoek, blijkt dat in de praktijk veel digitaliseringswerk wordt uitgevoerd met een collectie als uitgangspunt, in de stille hoop dat die, eenmaal digitaal beschikbaar, wel voor onderzoek gebruikt zal gaan

¹⁵ J.G.S.J. van Maarseveen en P.K. Doorn (red.), *Nederland een eeuw geleden geteld: een terugblik op de samenleving rond 1900* (Amsterdam, te verschijnen).

¹⁶ Onder andere in het kader van het project *Historische Databank Nederlandse Gemeenten*, waarin onderzoekers van de universiteiten van Nijmegen en Amsterdam, het CBS, het NIDI en het NIWI participeren. Het CBS heeft nog diverse tellingen ingevoerd. Financiering wordt gezocht om deze tellingen te ontsluiten. Het CBS zoekt ook naar mogelijkheden om de *Volkstelling van 1971* toegankelijk te maken.

worden.¹⁷ Het aanbod van digitale informatie zal in zekere mate de vraag stimuleren, maar toch denk ik dat het onderzoeksbelang bij de prioritering van digitaliseringsvoorstellen meer gewicht moet krijgen. Dat dit zal leiden tot een grotere heterogeniteit van digitale bronnen en tot 'lacunes' in elektronische producties, omdat keuzes moeten worden gemaakt vanuit een onderzoeksoptiek moet mijns inziens voor lief worden genomen.

Welling vroeg zich in zijn artikel af wie al die gegevens gaat invoeren. Ik denk dat het niet altijd efficiënt is om onderzoekers zelf hun bronnen te laten invoeren. AIO's en OIO's zijn wellicht betrekkelijk goedkope arbeidskrachten, maar het is een verspilling van de toch al schaarse onderzoekscapaciteit in de humaniora om grote hoeveelheden invoerwerk door onderzoekers te laten doen. Organisaties als het NIWI kunnen onderzoekers helpen bij het opzetten en organiseren van hun digitaliseringsprojecten. De uitvoering van het invoerwerk kan eveneens door gespecialiseerde instituten of door data-entrybedrijven gebeuren. Doel bij het digitaliseren lijkt mij om gegeven een bepaalde doelstelling en een bepaald budget de uitvoering zo efficiënt en goedkoop mogelijk te laten plaatsvinden.

Ooit stond ik bij een ouderwetse, doch gerenommeerde ijzerwarenhandel in Utrecht op mijn beurt te wachten. De klant voor mij begon omstandig uit te leggen dat hij op zoek was naar een bepaald soort piefje voor een ringetje waarmee een palletje was bevestigd aan een dingetje. De ondernemer legde geduldig zijn opschrijfboekje op zijn kalende hoofd en keek mij over de schouder van de klant meewarig aan en sprak in onvervalst Utrechts: 'Aach meneer, 't is me waat, d'r wordt wat aafgeklood in het laand.' Vervolgens deed hij een greep in één van de duizenden houten bakjes die de muur achter hem besloegen en haalde daaruit een ook voor mij ondefinieerbaar stukje gekromd metaal te voorschijn, dat precies bleek te zijn wat de klant zocht. De rekening bedroeg 15 cent.

Hoewel ik vrees dat het NIWI het niveau van dienstverlening van deze ijzerwarenhandel nooit zal kunnen evenaren, en zeker niet tegen die prijs, wil ik pleiten voor een professionele aanpak van historische digitaliseringsprojecten. Wanneer een aanzienlijke inspanning voor data-entry gepleegd moet worden loont het om een deugdelijk plan op te stellen, dat alle fasen van het project beslaat. Dit geldt zowel voor data-entry ten behoeve van onderzoek als ten behoeve van digitale bronnenpublicaties.

Waar het om gaat is dat de hele 'digitale keten' wordt overzien. Wanneer onderzoek het doel is, zal de onderzoeksvraag zeker van invloed zijn op de dataverzameling en de te maken keuzes en selecties. Belangrijk is dat deze goed verantwoord en beschreven worden, omdat alleen op die wijze ook latere onderzoekers kunnen beoordelen welke beslissingen zijn genomen. Bij elektronische bronnenpublicaties worden echter ook keuzes gemaakt. Hoewel ik – daartoe uitgenodigd – wel eens heb uitgerekend hoeveel het digitaliseren van een kilometer archiefmateriaal zou kosten¹⁸ en dat het scannen van het complete boeken- en tijdschriftenbezit van alle wetenschappelijke bibliotheken ter wereld ca. 10.000 arbeidsjaren zou kosten, lijkt mij de aanpak van een kip zonder kop bij een historisch digitaliseringsproject niet raadzaam.

¹⁷ *De computer en het alfa-onderzoek: advies van de commissie geesteswetenschappen over de toepassing van de informatietechnologie bij het onderzoek op het gebied van de geesteswetenschappen* (KNAW, juli 1997). *Een digitale bibliotheek voor de geesteswetenschappen: aanzet tot een programma voor investering in een landelijke kennisinfrastructuur voor geesteswetenschappen en cultuur*, samengesteld door Erik Viskil (Beleidsnota informatie- en communicatietechnologie van het gebiedsbestuur geesteswetenschappen van NOW, Den Haag, december 1999).

¹⁸ *Wetenschappelijk Technische Raad SURF, Alles uit de kast: op weg naar een nationaal investeringsprogramma digitale infrastructuur cultureel erfgoed*, samengesteld door W. Adriaans en anderen (Utrecht, 1998).

Tenslotte; de fraaiste vorm van digitalisering omvat een combinatie van *imaging* en *data-entry*, omdat hiermee het beste van twee werelden wordt geboden. Het moge echter duidelijk zijn dat deze vorm van digitalisering tevens de meest kostbare is. In het volkstellingenproject is gekozen voor een dergelijke combinatie, waarbij van een kwart van alle gepubliceerde volkstellingspagina's behalve images ook direct analyseerbare tabellen zijn gemaakt. De inspanning (en kosten) zijn echter omgekeerd evenredig: het scannen van de images en het maken van een ontsluiting daarop vormde slechts een beperkt deel van de totale projectkosten. Ik meen daarom dat wij verantwoorde keuzen hebben gemaakt in dit project, dat overigens in slechts ca. 2 jaar werd uitgevoerd. Uit Welling's kritiek blijkt vooral dat hij zich slecht geïnformeerd heeft, zowel over het volkstellingenproject als over de dataset met pondtolregisters.