21 oktober 2009

Peter Doorn, Kees Mandemakers, Henk Wals en Joris van Zundert

LEVENSLAB

Repository van data en tools op historisch-demografisch gebied

1. Introductie

Alfalab is van start gegaan als een samenwerkingsproject van vijf KNAW-instituten: DANS, Fryske Akademy, Huygens, Meertens en VKS. Het project concentreert zich volgens het projectplan op twee onderdelen: het Tekstlab (gericht op tekstuele bronnen) en het Ruimtelab (gericht op geodata). Beide elementen worden verbonden in een Alfalab-portaal. Kort na de start van het project toonde ook het IISG belangstelling om deel te nemen aan Alfalab. Uit besprekingen bleek dat de toetreding van het een zesde KNAW-partner zou betekenen dat er een derde "lab" aan de pilot zou worden toegevoegd: het Levenslab (gericht op bevolkingsdata, in het bijzonder betrekking hebbend op de levensloop). Dat zou een zeer waardevolle aanvulling betekenen, die de pilot niet alleen omvangrijker, maar ook evenwichtiger maakt door een groter gedeelte van de geesteswetenschappen bij Alfalab te betrekken, in het bijzonder de sociale geschiedenis en historische demografie. Het IISG is bereid substantieel te investeren in Alfalab. Ook DANS is bereid nog wat extra bij te dragen, terwijl in het bestaande Alfalab-budget ook nog wat verschoven kan worden. Niettemin willen we het KNAW-bestuur ook vragen nog een extra bijdrage te leveren. Het gaat om een bedrag van k€ 70. De totale aanpassing op de projectbegroting komt uit op € 163.400.

Het Levenslab werkt met historische bevolkingsgegevens. De laatste jaren zijn er tal van belangrijke bronnen met deze gegevens gedigitaliseerd. Hierbij gaat het zowel om gegevens op microniveau zoals die van de Historische Steekproef Nederland (HSN) als geaggregeerde gegevens op landelijk, provinciaal of gemeentelijk niveau. DANS werkt bijvoorbeeld al jaren samen met onder andere het CBS, het IISG en de Radboud Universiteit bij het toegankelijk maken van historische volkstellingen. Ook de Fryske Akademy heeft historisch demografisch materiaal over Friesland digitaal toegankelijk gemaakt.

Binnen Alfalab willen we op lange termijn streven naar een integratie van enerzijds het Nederlands historische datamateriaal op het gebied van levenslopen en anderzijds op geaggregeerde gegevens op demografisch gebied en daarbij reeds ontwikkelde tools. In de huidige pilotfase van Alfalab ligt bij het Levenslab de nadruk op het realiseren van een portaal voor historische tellingen en het exploreren (door proof of concept applicaties) van de mogelijkheden tot integratie. Een publieksvriendelijke en goed onderhoudbare toegang tot de data van de HSN dient daarbij als eerste concrete pilot. Hierbij zal het distilleren van datasets voor verder onderzoek sterk worden vereenvoudigd op basis van

een eerste integratiemodel voor historische levensdata op micro-, meso- en macroschaal (Intermediate Data Structure).

In het volgende zullen de verschillende aspecten nader worden uiteengezet.

2. Het levenslab

Op dit moment bestaan er diverse onafhankelijke websites op het gebied van de Nederlandse historische demografie Er bestaat echter geen centraal punt waarop een overzicht wordt geboden van het rijke materiaal dat in Nederland voorradig is. De belangrijkste websites zijn:

- www.volkstellingen.nl: biedt toegang tot alle tabellen van de gedigitaliseerde historische volkstellingen in Nederland tussen 1795 en 1971, alsmede tot de virtuele volkstelling van 2001.
- Hub for Aggregated Social History (HASH): HASH biedt portaalsoftware, waarin eenvoudig selecties uit bestaande databases (zoals de Historische Databank Nederlandse Gemeenten (HDNG), Volkstellingen en de Landbouwenquête 1889) worden gemaakt en gehomogeniseerd voor verdere verwerking, zoals visualisatie in een GIS.
- De CBS website bevat een grote hoeveelheid gedigitaliseerd materiaal, bestaande uit Jaarboeken, Jaarcijfers, Nationale rekeningen, Historische publicaties, Provinciale verslagen, Maandstatistiek bevolking. Ook het Statline-systeem bevat veel historisch materiaal (o.a. tijdreeksen).
- De Historische Databank Nederlandse Gemeenten bevat een grote hoeveelheid historische gemeentelijke gegevens uit diverse bronnen en instituten/universiteiten. Het is de bedoeling dat deze data worden ontsloten via HASH (zie hierboven).
- De Historische Steekproef Nederland (HSN)

Bij het ontsluiten van deze data zijn er momenteel twee belangrijke richtingen te onderscheiden. Bij de microdata wordt er gewerkt aan een platform waarin alle microdata kunnen worden ingelezen, de zogenaamde Intermediate Data Structure (IDS). Bij de data op geaggregeerde niveaus (zoals gemeente of regio) wordt in het HASH-project software gebruikt die uit verschillende tabelstructuren geharmoniseerde tabellen genereert, die eenvoudig koppelbaar zijn met Geografische Informatiesystemen (GIS). Beide benaderingen vullen elkaar aan. Binnen de IDS worden ook data van een geaggregeerd niveau opgeslagen en ook binnen GIS worden tools ontwikkeld om bijvoorbeeld individuele migratiepaden in kaart te brengen.

Door middel van de centrale toegang op de historische bevolkingsdata via Alfalab zullen onderzoekers uit binnen- maar ook uit buitenland niet alleen een snel overzicht krijgen van het beschikbare materiaal, maar ook van de tools/webservices die daarop van toepassing zijn.

3. De HSN toegankelijk gemaakt via het Levenslab

Het eerste pilotproject van het Levenslab richt zich op de HSN. De HSN is een internationaal gerenommeerde database gebaseerd op een *at random* getrokken steekproef uit de geboorteregisters van de periode 1812-1922. De kwaliteit en de opzet van het Nederlandse bevolkingsregister maken het mogelijk de levenslopen van onderzochte personen te volgen, ook als ze verhuizen, en daarmee levenslopen te reconstrueren. Dit maakt de HSN internationaal tot een unieke database. Daarnaast zijn er voor diverse onderzoeken kleinere bestanden aangelegd van *oversamplings* met hele of gedeeltelijke levenslopen. Van ongeveer 40.000 onderzoekspersonen zijn de volledige levenslopen inmiddels gereconstrueerd en in een dynamische database opgenomen.

Gegeven de bovengenoemde problematiek van het bronnenmateriaal is het programma voor de verwerking van de ingevoerde registers tot levenslopen een tamelijk complex geheel. De levenslopen zijn gebaseerd op soms wel tientallen inschrijvingen per persoon uit het bevolkingsregister over de periode 1850 tot heden. Het aantal inschrijvingen is afhankelijk van de periode waarin het register is opgemaakt, het aantal verhuizingen en de levensduur van de onderzoekspersoon. Sinds 1850 zijn er vijf systemen gehanteerd bij het register en in het bijzonder in de periode 1850-1900 hielden gemeenten er ook eigen gewoonten op na bij de wijze waarop het register werd bijgehouden en personen werden ingeschreven. Daarnaast werden er fouten gemaakt door de ambtenaren, waren ze onvolledig bij het inschrijven, of hielden ze zich niet aan de procedures. Ook hielden personen zich lang niet altijd aan de plicht zich op tijd en volledig te laten inschrijven. Ten slotte zijn er hier en daar ook registers verdwenen, vooral als gevolg van oorlogshandelingen tijdens de Tweede Wereldoorlog. Het grootste probleem is het veelal ontbreken van goede dateringen in de registers. Dit betekent dat dateringen van de ene gebeurtenis moeten worden afgeleid uit een andere of een serie andere gebeurtenissen die exact of in elk geval preciezer zijn gedateerd.

De bovengeschetste complexiteit brengt met zich mee dat het geen simpele zaak is om een voor onderzoekers gemakkelijk te hanteren output te genereren.

De situatie van de HSN is niet uniek. Er bestaan in de wereld vergelijkbare databases, zoals de Demographic Database Umeå en de Utah Population Database (Salt Lake City), gebaseerd op bevolkingsregistraties of vergelijkbare bronnen, die eveneens levenslopen produceren.[1] Al deze grote databases kampen met het probleem dat zowel de data als de onderzoeksvragen tamelijk complex zijn. Voor de meeste databases is het daarom problematisch om met algemene releases te komen waaruit onderzoekers kunnen putten. Bij de Demographic Database in Umeå bijvoorbeeld zijn er speciale programmeurs die met onderzoekers de vragen doornemen en vervolgens een voor het desbetreffende onderzoek specifieke dataset maken. Maar dit is een kostbare situatie en bovendien ook weinig flexibel (onderzoekers moeten naar Umeå komen voor overleg etc.).

In 2006 nam de HSN de leiding in de internationale discussie hoe tot een oplossing te komen voor deze problematiek. Inmiddels zijn er drie workshops geweest, mede

[1] Voor een overzicht van historische databases op basis van bevolkingsregisters en vergelijkbaar materiaal, zie http://historicaldemography.net/questionnaires.php.

gefinancierd door NWO-Geesteswetenschappen, en is er een consensus over de wijze waarop deze problematiek moet worden opgelost.[2] In essentie komt het erop neer dat de data van de verschillende databases op een tamelijk basaal niveau in een gemeenschappelijke datastructuur worden ondergebracht, de zogenaamde *Intermediate Data Structure* (IDS). Op basis van deze IDS worden er applicaties ontwikkeld die de data in een structuur omzetten die voor een bepaald onderzoek gewenst is. Zowel IDS als de onderzoeksspecifieke software worden gedeeld en verspreid door middel van een collaboratory. Er is inmiddels een werkend voorbeeld van een tool om een dataset te bouwen voor onderzoek naar vruchtbaarheid op basis van de IDS (ontwikkeld door George Alter, Interuniversity Consortium for Political and Social Research (ICPSR)). Bij verschillende databases is men inmiddels begonnen de data te converteren naar de IDS (onder andere bij DDB Umeå en bij de Scania database in Lund).

Om de levensloopdata via Alfalab beter toegankelijk te maken voor onderzoekers is het nodig dat de HSN-gegevens in de IDS-structuur worden omgezet. Dit betekent dat de daarop gebaseerde tools betrekkelijk simpel en goedkoop kunnen zijn. De bouw van de IDS omvat de volgende elementen:

1. Standaardiseren HSN-gegevens naar de IDS-structuur
2. Identificeren en linken van alle vermeldingen die in de diverse inschrijvingen voorkomen maar tot één unieke onderzoekspersoon behoren.
3. Geven van dateringen daar waar bepaalde gebeurtenissen of kenmerken van personen niet, slecht of inconsistent zijn gedateerd. Dit komt in het bronnenmateriaal onder meer voor bij verhuizingen, beroepstitels en vermeldingen van godsdienst.
4. Bouwen van een semi-automatische structuur om optredende inconsistenties bij de onderdelen 2 en 3 op te lossen. Deze structuur bestaat uit software om a) fouten en inconsistenties te traceren en te melden en b) waar mogelijk tot automatische oplossingen te komen.

Op basis van het voorwerk door middel van de IDS kan als pilot vervolgens op een vrij gemakkelijke en goedkope wijze een datatool voor geografische mobiliteit (migratie) worden gebouwd. Dit zouden we willen realiseren in de eerste fase van Alfalab. Het gaat om een dataset die standaard de elementen wegzet die een doorsnee onderzoeker op dit gebied nodig heeft. Dit zijn in elk geval alle combinaties van kenmerken van het huishouden, woonplaatsen en uitgeoefende beroepen. Alle combinaties worden van een datum voorzien, zodat het mogelijk is event history analyse toe te passen. Bij de kenmerken van het huishouden moet gedacht worden aan zaken als structuurtypering (eenvoudig gezin of uitgebreid met grootouders, of neven, nichten) en grootte van het

[2] 'Towards a global history of life courses. Creating a network for the development of data structures for standardized longitudinal historical data', granted by ICPSR Ann Arbor, DDB Umeå and Netherlands Organisation for Scientific Research (NWO), Humanities (Internationalizing, 236-53-004). Een beschrijving van de opzet zal verschijnen als George Alter, Kees Mandemakers and Myron Gutmann, 'Defining and Distributing Longitudinal Historical Data in a General Way through an Intermediate Structure', *Historical Social Research*, to be published August 2009. Een samenvatting in K. Mandemakers, *Waarom Jan en Cor trouwden. Over grote historische databestanden, koudwatervrees en interdisciplinaire samenwerking*. Oratie Erasmus Universiteit Rotterdam, 11 juni 2009 (Amsterdam: Aksant 2009), 31-33.

huishouden. Het overnemen of schatten van een datum bij alle gegevens is een zeer lastige zaak die voor de onderzoeker in de IDS al is opgelost. De bouw van de migratietool vindt plaats in overleg met onderzoekers op dit gebied. Een individuele onderzoeker die op bepaalde punten meer wil, kan zelf op basis van de IDS zijn dataset op relatief gemakkelijke wijze aanvullen, in principe wordt de daarvoor gebouwde tool weer aan de verzameling toegevoegd.

4. Integratie in het Alfalab-portaal

Het Levenslab biedt DANS de gelegenheid om in het kader van Alfalab een repository op te zetten, waarin de data en webservices van deze projecten vanuit één plek kunnen worden gevonden. De Nederlandse data zullen daarbij zoveel mogelijk in DANS EASY worden gearchiveerd. Het portaal biedt niet alleen een overzicht van en zoekmogelijkheid naar de beschikbare hulpbronnen maar wordt a) ook een ingang op door de onderzoeksgemeenschap ontwikkelde tools die de data ontsluiten, en b) op internationaal belangrijke websites met historisch demografisch materiaal, zoals die van IPUMS (Integrated Public Use Microdata Series) van het Minnesota Population Centre.

5. De gebruiker van het levenslab

De wijze waarop onderzoekers het levenslab zullen gebruiken zal sterk afhankelijk zijn van het doel en de ervaring van de gebruikers. De volgende voorbeelden mogen een kleine impressie geven:

1 Een gebruiker krijgt via het Levenslab toegang tot de HSN en kan met behulp van één van de in het Levenslab aanwezige selectie- en migratietools een dataset distilleren uit de data die zijn opgeslagen in de IDS . Vervolgens kan hij door middel van de ingang op de HASH-database geaggregeerde gegevens op gemeentelijk niveau binnenhalen.

2 Een gebruiker komt via het Levenslab bij de collaboratory ‘Historical Life Courses’ en krijgt via de collaboratory toegang tot de plaats waar de tools worden gedocumenteerd en bewaard. Hier kan hij zelfgemaakte tools laten registreren en op deze wijze beschikbaar stellen aan de onderzoekgemeenschap. Via deze collaboratory kan hij ook vinden welke bestanden er wereldwijd allemaal beschikbaar zijn waarbij per onderzoeksgebied wordt aangegeven welke data geschikt zijn en welke selecties gemaakt kunnen worden (via deze ingang is dus ook de HSN te benaderen).

3 Een onderzoeker vindt via het Levenslab een overzicht van alle relevante databronnen met betrekking tot de Nederlandse bevolking op onderwerp, plaats en jaar, en van de belangrijkste vergelijkbare bronnen in het buitenland. Hij kan eenvoudig selecties maken en krijgt als resultaat een bestand dat gemakkelijk verder te verwerken is voor statistische analyse of visualisatie in een GIS of grafiekprogramma.

6. Beoogde samenwerking

De aan de Intermediate Data Structure (IDS) gekoppelde tool voor onderzoek naar migratie met de daarin opgenomen geografische coördinaten sluit aan op het Ruimtelab. De bouw van een IDS biedt ook een oplossing voor een betere (her)benutting van andere in Nederland bestaande datasets met levensloopgegevens. Met behulp van de op basis van dit plan ontwikkelde software zullen namelijk ook andere datasets betrekkelijk gemakkelijk in een IDS-structuur kunnen worden omgezet. De tool voor geografische mobiliteit sluit ook aan op de bestaande initiatieven waarbij op gemeentelijke en provinciale niveaus contextuele gegevens worden verzameld en voor onderzoek ter beschikking worden gesteld (HASH, DANS: volkstellingen). Binnen de KNAW betekent dit vooral een samenwerking tussen de Fryske Akademy, HSN/IISG, NIDI en DANS, buiten de KNAW met de universiteiten, met name Radboud Universiteit (HASH) en met internationale data-instituten als de Demographic Database Umeå, IPUMS en ICPSR.

De onderzoeksgemeenschap zal vanaf het begin bij het Levenslab worden betrokken. Aan het eind van de pilotfase wordt een workshop georganiseerd waarin de behaalde resultaten worden gepresenteerd en bediscussieerd en waarin de contouren voor een verdere uitbouw worden vastgesteld.

Kosten bouw IDS, migratie-tool en repository

	Uren	Tarief	Kosten
IDS			
Functioneel ontwerp	500	52	26.000
Implementatie IDS	100	52	5.200
Ontwikkeling programma	1400	52	72.800
Testen uitkomsten programma	500	52	26.000
Subtotaal IDS			130.000
Bouw migratie-tool			
Functioneel ontwerp	100	52	5.200
Ontwikkeling programma	250	52	13.000
Testen uitkomsten programma	100	52	5.200
Subtotaal migratie-tool			23.400
Repository levenslab			10.000
Algemeen totaal			163.400

Voorgestelde dekking

IISG	76.700
Alfalab	6.700
DANS	10.000
KNAW	70.000
Algemeen totaal	163.400