



Centraal Bureau voor de Statistiek

Procesontwikkeling, IT en Methodologie

Procesontwikkeling en Methodologie

DE INZET VAN INTERNETWAARNEMING BIJ DE VACATURESTATISTIEKEN; EINDRAPPORTAGE WERKZAAMHEDEN 2011

Nico Heerschap, Jacco Daalmans, Dick ter Steege

Samenvatting: Vanaf 2010 wordt verkennend onderzoek verricht naar de toepassingsmogelijkheden van internetdata voor de Vacaturestatistieken. Er zijn analyses uitgevoerd op bestanden, die verkregen zijn van het bedrijf Textkernel. Dit bedrijf is gespecialiseerd in het waarnemen, structureren en classificeren van op het Internet gepubliceerde vacatures. Doelstellingen van het onderzoek zijn geweest: 1) het vervangen van de waarneming van de Structuurenquête Vacatures door internetwaarneming en 2) het beperken van de uitvraag van de Kwartaalstatistiek Vacatures door het benutten van internetwaarneming. In deze nota worden de bevindingen van de eerste verkennende analyses gepresenteerd, die in 2010 en 2011 zijn uitgevoerd.

De hoofdconclusie is dat op basis van de tot nu toe uitgevoerde analyses de gestelde doelstellingen vooralsnog niet kunnen worden verwezenlijkt. Wel zijn er aanknopingspunten voor verder onderzoek, mede ook omdat de beschikbare data tot op heden relatief beperkt waren.

Er is wel een relatie geconstateerd tussen bedrijven die bij de Kwartaalstatistiek opgeven geen vacatures te hebben en waar Jobfeed ook geen vacatures op het Internet waarneemt en vice versa. Als naar die bedrijven bijvoorbeeld in drie kwartalen geen vragenlijst zou worden uitgestuurd, zou dat al een aanzienlijke beperking van de enquêtedrukvermindering opleveren. Ook hier is echter meer onderzoek nodig.

Trefwoorden: Jobfeed, internetrobots, waarneming op het internet, vacatures, lastendruk verlichting

Projectnummer:

209503

BPA-nummer:

PPM-2012-2-22-NHRP-JDAS-DSGE

Datum:

22 februari 2012

1. Inleiding

Het Internet is een fundamentele infrastructuur geworden van de samenleving. Steeds meer activiteiten uit de “fysieke” wereld verplaatsen zich naar het Internet. Daarmee is het Internet niet alleen een fenomeen geworden dat vraagt om een statistische beschrijving, maar ook een nieuwe bron van informatie, die voor het maken van statistieken benut kan worden. Dit is mogelijk omdat activiteiten op het Internet digitale sporen, *digital footprints*, achterlaten. Deze digitale sporen van miljoenen mensen kunnen 24 uur per dag en 7 dagen per week worden waargenomen.

Ook als het gaat om het werven van personeel speelt Internet een belangrijke rol (Intermediair, 2011). De meeste bedrijven hebben tegenwoordig wel een website, waarop zij informatie verstrekken over de bij hen beschikbare vacatures. Daarnaast zijn er intermediairen (job boards), die internet services aanbieden waarop bedrijven hun vacatures kunnen plaatsen. Voorbeelden zijn Monsterboard.nl en Werk.nl. Hoewel nog beperkt van omvang is een nieuwe ontwikkeling het gebruik van sociale media voor het werven van personeel. Een bekende service op dit terrein is LinkedIn, maar de werving kan ook gaan via bijvoorbeeld Facebook, Hyves of Twitter.

Het verzamelen van data van het Internet kan op verschillende manieren (Dialogic, 2008). Eén van de mogelijkheden is het inzetten van zogenaamde internetrobots – crawlers of -spiders. Het betreft software, waarmee het Internet kan worden afgezocht naar de gewenste informatie. Daarbij kan het gaan om het benaderen van één of meer specifieke websites, maar het kan ook gaan om een scan van een deel van het Internet (web scraping). Bijvoorbeeld het Nederlandstalige deel van het Internet. De opgezochte informatie wordt centraal opgeslagen in een database, die vervolgens de basis vormt voor het maken van statistieken. Doordat bij deze methode de informatie op websites wordt doorzocht, wordt hier gesproken over een *website-centric* benadering¹.

Onderzoek naar de toepassingsmogelijkheden van internetrobots voor het maken van statistieken loopt al sinds 2009 bij het CBS. In eerste instantie ging het om prijsinformatie van benzine en vliegreizen (Hoekstra e.a., 2010). Vanaf 2010 zijn daar onder de paraplu van het speerpunt Impact ICT (Heerschap, 2010B en Heerschap, 2011) de domeinen woningmarkt en vacatures aan toegevoegd. Het domein van vacatures is vooral in beeld gekomen, doordat steeds meer partijen, zoals bijvoorbeeld Monsterboard en Textkernel, statistieken over het aantal en de ontwikkeling van vacatures publiceren die gebaseerd zijn op internetwaarneming. Publicaties die de aandacht trekken van de beleidsmakers.

¹ Andere benaderingen zijn: 1) *user-centric*, waarbij bijvoorbeeld software wordt geïnstalleerd op de pc of smartphone van iemand; deze software registreert op de achtergrond het gebruik; en 2) *network-centric*, waarbij gemeten wordt op knooppunten in het Internet, bijvoorbeeld hoeveel dataverkeer langs het knooppunt komt.

In het kader van onderzoek naar de toepassingsmogelijkheden van internetwaarneming voor de vacaturestatistieken heeft het CBS in 2010 en 2011 de beschikking gekregen over data uit de vacature-database Jobfeed van Textkernel. Deze nota beschrijft de bevindingen van dit onderzoek en geeft in algemene termen aan welke richtingen kansrijk zijn voor verder onderzoek.

De resultaten van die werkzaamheden zijn beschreven in de hoofdstukken 4 en 5. Alvorens daarop in te gaan worden, in hoofdstuk 2, in het kort, de potentiële voordelen van het gebruik van data van het Internet voor het maken van statistieken beschreven. In hoofdstuk 3 wordt een korte schets gegeven van de vacaturestatistieken bij het CBS en de relatie Internet en vacatures. De notitie wordt in hoofdstuk 6 afgesloten met de belangrijkste conclusies en mogelijke vervolgstappen. In bijlage I wordt wat meer achtergrond verstrekt over de vacaturedatabase Jobfeed.

2. Voordelen van Internet als databron

Het gebruik van Internet als nieuwe databron voor het maken van (officiële) statistieken kan, in potentie, onder meer de volgende voordelen opleveren:

- *Nieuwe statistieken.* De hoeveelheid data die op het Internet kan worden waargenomen is niet meer beperkt door de omvang van het enquêteformulier of de verwachte enquêtedruk. Zolang de data op het Internet beschikbaar zijn, kunnen deze worden waargenomen, ongeacht de hoeveelheid variabelen en de hoeveelheid data. Dit biedt mogelijkheden om nieuwe statistieken te maken, die met traditionele wijze van waarnemen niet of moeizaam te realiseren zijn. Daarbij speelt ook de verschuiving van het achteraf vragen stellen aan de respondent naar het real-time observeren van het gedrag van de respondent, hier vastgelegd in digitale sporen op het Internet.
- *Snellere statistieken.* Er hoeft niet meer gewacht te worden totdat een enquête is uitgezet, teruggekregen en verwerkt. Data, c.q. de digitale sporen, van het Internet kunnen 24 uur per dag en 7 dagen per week worden waargenomen. Er is, althans voor wat de waarnemingsfase betreft, sprake van een nagenoeg real-time verzameling van de data. Bij gelijkblijvende tijd voor analyse, kan dan sneller worden gepubliceerd. Snellere cijfers betekent onder meer dat nieuwe ontwikkelingen, zoals bijvoorbeeld een omslag in het aantal vacatures, sneller zichtbaar worden.
- *Meer detail en een betere kwaliteit.* Doordat op het Internet vaak grote hoeveelheden data over een onderwerp beschikbaar zijn, kan het gemakkelijker worden om op een meer gedetailleerd niveau te publiceren dan nu mogelijk is (bijvoorbeeld regionaal). Door de van het Internet verkregen data te combineren

met bestaande statistieken is het ook mogelijk de kwaliteit van die statistieken te verhogen.

- *Enquêtedrukvermindering.* Informatie hoeft niet meer aan respondenten, bijvoorbeeld bedrijven, te worden gevraagd, maar wordt waargenomen op het Internet zonder daarvoor de respondenten lastig te vallen.
- *Lagere operationele kosten.* Het CBS hoeft zelf geen dure enquêtes meer uit te voeren, maar kan direct waarnemen via het Internet.

Bovenstaande voordelen sluiten naadloos aan bij de strategische doelen van het CBS. Snellere en nieuwe statistieken die aansluiten bij de wensen van haar klanten, voor lagere kosten waarbij geen sprake is van enquêtedruk.

Het gebruik van het Internet als databron voor het maken van statistieken is buiten de statistische bureaus al lang gemeengoed geworden. Hoewel men vaak vraagtekens kan zetten bij de kwaliteit van dit soort statistieken, trekt deze nieuwe manier om cijfers te maken veel aandacht, met name ook van beleidsmakers. In dit kader wordt het CBS vaak als "ouderwets en traag" gekwalificeerd (zie bijvoorbeeld Binnenlands Bestuur, 2012). Het CBS kan bij dit soort nieuwe ontwikkelingen dan ook niet langer aan de zijlijn blijven staan en zal moeten onderzoeken of dit soort technieken ook voor het maken van officiële statistieken nut heeft. Het hier gepresenteerde onderzoek van internetwaarneming voor het samenstellen van vacaturestatistieken is één van die toepassingsmogelijkheden.

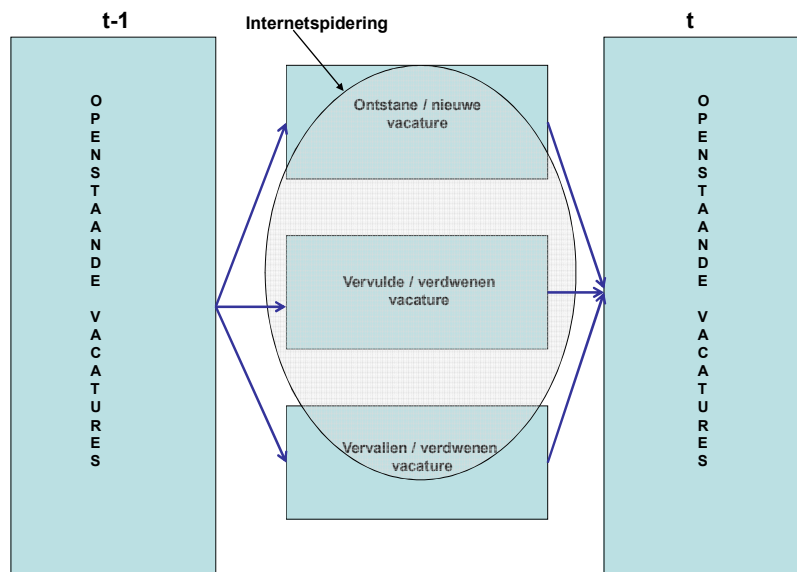
3. De vacaturestatistieken en het Internet

3.1 De vacaturestatistieken

Bij de vacaturestatistieken gaat het om:

1. *De Kwartaalstatistiek Vacatures.* Bij deze statistiek wordt gevraagd naar de openstaande vacatures aan het eind van een kwartaal en gedurende het kwartaal ontstane en vervulde vacatures. Het CBS houdt deze enquête bij circa 22 duizend bedrijven en instellingen in zowel de particuliere- als in de overheidssector, inclusief het onderwijs. Doel is om uiterlijk 45 dagen na afsluiting van een kwartaal cijfers te publiceren over de ontwikkeling van de vacaturemarkt, w.o. het aantal openstaande vacatures naar sector (SBI-groep). De statistiek is gebaseerd op een Europese verordening.

Figuur 1. Kader vacatures



De waarneming vindt plaats door formulieren naar bedrijven te sturen. Hierbij wordt naast het aantal vacatures ook gevraagd naar het ziekteverzuim. Dit veroorzaakt een aanzienlijke enquêtedruk.

Opvallend is dat er een behoorlijk verschil kan zijn tussen het totaal door bedrijven opgegeven vacatures (ruwe data) en het uiteindelijke aantal vacatures na analyse (de gaafgemaakte data, zie tabel 7 in hoofdstuk 5). Dit komt mede omdat bij de analyse de volgende rekenregel wordt gehanteerd:

$$\text{Openstaande vacatures } t = \text{Openstaande vacatures } t-1 + \text{nieuwe vacatures} - (\text{vervulde en vervallen vacatures}),$$

zie figuur 1. Indien de ruwe gegevens niet aan deze regel voldoen is dat reden voor aanpassing.

2. *De Structuurenquête Vacatures*. De vragen uit de Kwartaalstatistiek zeggen niets over wat voor soort vacatures er zijn. Om hierachter te komen, wordt de vraag naar vacatures eens in de twee jaar uitgebreid met een extra vragenlijst (3de kwartaal). Deze lijst wordt de Structuurenquête genoemd. Daarbij gaat het tot nu toe om de volgende kenmerken:

- Gemeente waar de vacatures openstaan;
- Functie;
- Aantal medewerkers aan wie leiding moet worden gegeven;

- Vereiste opleiding en vakrichting;
- Aantal vacatures waarvoor geen werkervaring is vereist;
- Aantal vacatures waarvoor korter dan één maand wordt gezocht;
- Aantal vacatures waarvoor tussen de één en drie maanden wordt gezocht;
- Aantal vacatures waarvoor een personeelsadvertentie is geplaatst;
- Aantal vacatures dat is aangemeld bij het Centrum voor Werk en Inkomen;
- Aantal vacatures dat een bedrijf als moeilijk vervulbaar beschouwt;
- Aantal vacatures dat alleen is bestemd voor interne kandidaten;

De Structuurenquête is gebaseerd op een gentleman's agreement met de Europese Unie. Met name bij het UWV is behoefte aan de informatie uit deze statistiek. De laatste Structuurenquête is in 2008 geweest. In 2010 is de statistiek uitgesteld om te bezien of er alternatieven waren voor de papieren uitvraag, die administratieve lasten met zich meebrengt². Als één van de alternatieven wordt het verzamelen van data van het Internet over vacatures gezien. Daarbij zou het aantal variabelen wel drastisch worden beperkt.

3. *Vacature indicatoren* Ten slotte, wordt er bij de NR (conjunctuur) een vacature-indicator gepubliceerd die maandelijks indicaties geeft over de ontwikkelingen in de vacaturemarkt. Deze indicator is afgeleid uit verschillende variabelen van de maandelijks conjunctuuronderzoeken die door het Centraal Bureau voor de Statistiek en het Economisch Instituut voor de Bouw worden gehouden. De vacature-indicator geeft een kwalitatieve indicatie van de richting waarin de vacatures zich naar verwachting van de ondernemers zullen ontwikkelen, zie Ruth en Wekker (2009 en 2011).

3.2 Internet en vacatures

Het werven van personeel via het adverteren met of het *posten* van vacatures heeft altijd al op het Internet plaatsgevonden. De meeste bedrijven, die op het Internet vertegenwoordigd zijn, plaatsen de beschikbare vacatures op hun website. Een belangrijke ontwikkeling aan het eind van de vorige eeuw is de opkomst van intermediairs of jobboards geweest. Deze jobboards bieden specifieke vacaturesites aan, waarop bedrijven hun vacatures tegen betaling kunnen plaatsen. Het aantal van dit soort sites in Nederland is in de loop van tijd sterk gegroeid, namelijk tegen de

² De Structuurenquête telt echter niet mee in de officiële berekening van de enquêtedruk, omdat er geen sprake is van een verplichte statistiek.

600 (Intermediair, 2011 en Drees, 2012A). De belangrijkste zijn o.a. Werk.nl (van het UWV), Monsterboard.nl, NationaleVacaturebank.nl, Jobtrack.nl, Stepstone.nl en Intermediair.nl. Naast deze algemene vacaturesites, zijn er ook kranten, die hun vacature-aanbod op het Internet plaatsen en sites die zich op specifieke doelgroepen richten, zoals JobsInFinance en Villamedia (journalistiek). Door het grote aanbod van vacatures op verschillende plekken zijn er ook services (aggregate-sites) ontstaan die voor de werkzoekende in één keer alle of veel vacature-sites doorzoeken. Voorbeelden zijn AskJim.nl, CareerJet.nl en JobRapido.nl. Met de snelle opkomst van de sociale media in het begin van deze eeuw is het niet verbazingwekkend dat ook via deze kanalen steeds vaker personeel wordt geworven. LinkedIn is het bekendste voorbeeld, maar er wordt ook steeds meer gebruik gemaakt van de bekende sociale media zoals Facebook, Hyves en Twitter. Een nieuwe ontwikkeling van de laatste paar jaar is de toename en het belang van gratis vacaturesites, zoals Banenmatch. De laatste twee ontwikkelingen, sociale media en gratis vacaturesites, hebben er toe bijgedragen dat de traditionele job boards het moeilijk hebben. Ten slotte zijn er organisaties die het Internet afzoeken naar vacatures en de informatie daarvan opslaan in een database met variabelen zoals postcode, opleiding, beroep, url e.d. De opgeslagen informatie wordt aangeboden voor gebruik door derde partijen, met name om marktonderzoek te doen en statistieken te maken. Workinglife.nl en Textkernel (Jobfeed) zijn dergelijke bedrijven. In dit onderzoek is gebruik gemaakt van de gegevens van Textkernel.

Al deze ontwikkelingen zijn niet onbelangrijk als het gaat om het maken van statistieken over vacatures op basis van internetwaarneming. Aspecten die spelen bij het waarnemen van vacatures op het Internet:

- De groei van het aantal op het Internet waargenomen vacatures kan (mede) het gevolg zijn van het toenemende gebruik door bedrijven, instellingen en overheid van het Internet als wervingskanaal. Het daadwerkelijke aantal vacatures hoeft in feite niet te zijn gestegen.

Tabel 1. Marktbereik wervingskanalen, 2006 (bron Grijpstra e.a., 2010).

Sector	In %	Advertentie	Uitzendbureau	Internet
Agrarisch		36	17	4
Industrie		37	31	26
Bouw		30	28	29
Handel, horeca		32	7	22
Zakelijke dienstverlening		37	14	51
Overheid		49	11	58
Overig		46	10	33
Totaal		36	14	38

- Een aanzienlijk deel van alle beschikbare vacatures bij bedrijven, instellingen en overheid wordt nog niet op het Internet geplaatst. Dat verschilt sterk per SBI-groep. Zie tabel 1, waar het marktbereik van 2006 van de verschillende

wervingskanalen is aangegeven: het totaal aantal vacatures waarvoor een wervingskanaal is ingeschakeld ten opzichte van alle vacatures. Recentere cijfers zijn niet beschikbaar. In Grijpstra e.a. (2010) wordt de indruk gewekt dat het markt bereik (totaal) van het Internet in 2009 rond de 50% ligt. Een stijging van 12% in drie jaar.

- Een essentieel probleem bij het vaststellen van het aantal unieke vacatures op het Internet is het detecteren en verwijderen van vacatures die op meer dan één plek worden gepost. Dit proces van ontdebellen is geen sinecure. Niet is het alleen van belang hoe uit de tekst van de vacature wordt geconstateerd dat er sprake is van een dubbele vacature, ook is het van belang hoe ver men teruggaat met deze controle. Zo wordt bij de vacature database Jobfeed tot op heden 3 weken teruggedaan om vacature-teksten met elkaar te vergelijken. Dat lijkt te kort. Men is nu aan het experimenteren met een periode van 6 weken. Een extra complicatie vormt het groeiende belang van gratis vacaturesites en sociale media, waardoor dezelfde vacature op nog meer plekken op het Internet te vinden is.
- Een ander probleem betreft de zogenaamde bulkvacatures, waarbij in één advertentie wordt gevraagd naar bijvoorbeeld 3 programmeurs. Het is niet gemakkelijk uit de tekst te halen dat er in feite sprake is van 3 in plaats van 1 vacature. De indruk is dat het slechts gaat om een relatief klein aantal vacatures (Grijpstra e.a., 2010). Onderzocht is het niet.
- Als vacatures worden waargenomen bij intermediairs, uitzendbureaus e.d. is er meestal te weinig informatie beschikbaar, zoals de naam van de werkgever, het adres e.d., om de vacature goed te identificeren. Koppeling met bijvoorbeeld Kamer van Koophandelnummer (KvK-nummer) of classificatie naar regio en SBI is door het ontbreken van gegevens vaak niet mogelijk. Dat maakt het ook moeilijk om deze vacatures te betrekken bij het ontdebellen en het uiteindelijk maken van statistieken. Het gaat hier om grote aantallen (Grijpstra e.a., 2010).
- Verder speelt het bereik en de techniek van de spidering een rol. Worden alle vacatures ook daadwerkelijk gevonden op het Internet? Ook dit proces wordt steeds beter, met als gevolg een toename van het aantal vacatures, zonder dat er daadwerkelijk sprake hoeft te zijn van een toename.
- Het aantal waar te nemen vacatures kan, ten slotte, worden beïnvloed door strategieën van de job boards. Zo is het mogelijk dat vacatures langer op het Internet blijven staan. Het aantal vacatures dat een job board heeft, bepaalt namelijk ook de plek die zo'n service inneemt op de lijst met belangrijkste vacaturesites (Drees, 2012B). Ook worden er termijnen gehanteerd dat een vacature op het Internet mag staan, bijvoorbeeld een maand of zes weken (zie Van Velzen e.a., 2011). Door de bovenstaande problemen is het lastig om af te leiden wanneer een internet vacature is vervuld of verdwenen.

- Bedrijven, die een vacature niet kunnen vervullen, halen deze soms weg en plaatsen de vacature op een later tijdstip met een iets gewijzigde tekst opnieuw. De vraag is dan of er sprake van een nieuwe vacature? Hier ligt ook een relatie met de periode die men teruggaat om op dubbele vacatures te controleren.
- Het onderscheid tussen vervallen en vervulde vacatures kan via internetwaarneming niet worden vastgesteld. Daar kan alleen worden gezien of een vacature niet meer op het Internet staat, dus verdwenen is. Openstaande vacatures worden bij internetwaarneming vastgesteld op basis van het feit dat ontstane vacatures nog niet van het Internet verdwenen zijn. Vacatures worden als ontstaan gekwalificeerd op het moment dat zij op het Internet zijn gevonden. Dus niet op het moment dat zij geplaatst zijn.

Het is in de praktijk moeilijk gebleken om al dit soort effecten te kwantificeren en deze als parameters mee te nemen bij de analyse en in een later stadium bij de ophoging naar representatieve aantallen.

3.3 Kamer van Koophandelnummers

Van groot belang voor het onderzoek naar de toepassingsmogelijkheden van het gebruik van internetwaarneming voor statistieken is de mogelijkheid de op het Internet gevonden vacatures te kunnen koppelen aan en te kunnen vergelijken met de waarneming van de Kwartaalstatistiek en Structuurenquête. Daarvoor is het nodig dat de gevonden vacature wordt voorzien van een Kamer van Koophandelnummer (KvK-nummer). Bij de vacaturedatabase Jobfeed gebruikt men een module met algoritmes die naam en adresgegevens van bedrijven koppelt aan KvK-nummers. Is een internetvacature voorzien van een KvK-nummer dan kan deze via een koppeltabel worden gekoppeld aan de Bedrijfseenheid, de statistische eenheid van de vacaturestatistieken. De op deze wijze gekoppelde data biedt veel meer mogelijkheden voor onderzoek dan vacaturedata zonder KvK-nummers.

4. Uitgevoerde werkzaamheden

4.1 Achtergrond

In 2009 is bij het CBS gestart met kleinschalig onderzoek naar de toepassingsmogelijkheden van het Internet als nieuwe databron. Daarbij ging het om gegevens van het Internet, die door een internetrobot zijn verzameld. De domeinen betroffen prijzen van benzine en vliegreizen (Hoekstra e.a., 2010). In de loop van 2010 is dat uitgebreid met de woningmarkt en vacatures.

Een schematische weergave van het proces van het samenstellen van statistieken op basis van internetwaarneming staat in figuur 2.

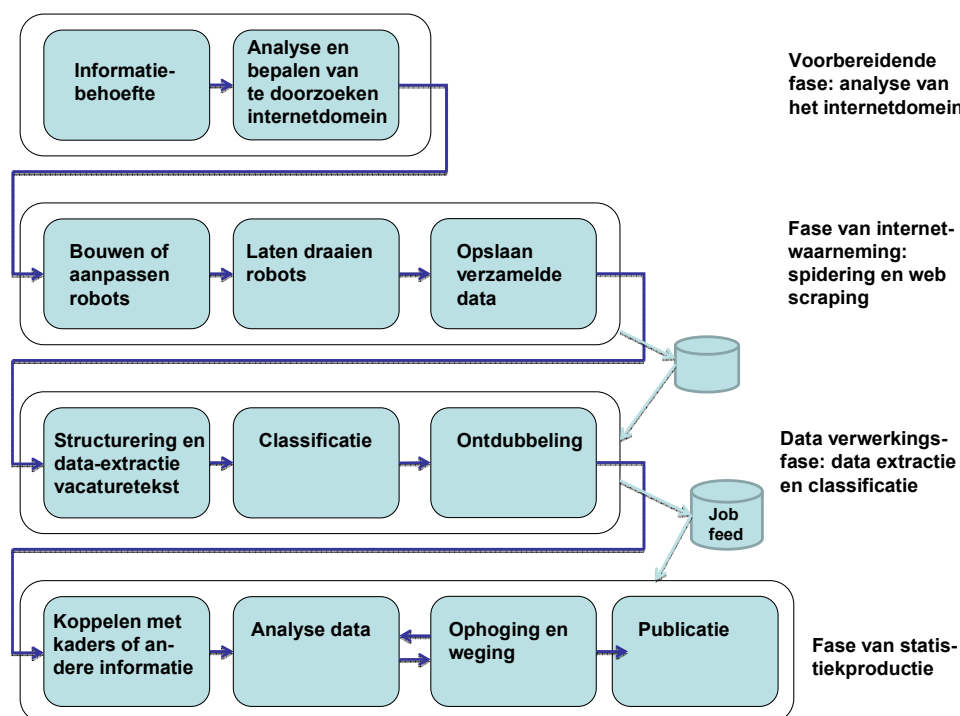
Bij de start van het project met internetvacatures was het hoofddoel het vervangen van de papieren waarneming van de tweejaarlijkse Structuurenquête door internetwaarneming, echter wel met een aanzienlijk beperktere set van variabelen. Het eerste idee was het zelf bouwen van eigen internetrobots, gericht op de bekendste vacaturesites en de websites van de top 200 bedrijven (Heerschap, 2010A).

Al snel kwam echter het UWV in beeld, die al een jaar bezig waren met het experimenteren met internetdata voor het aanbieden van vacatures aan werkzoekenden. Ook was opdracht gegeven aan het EIM en Research voor Beleid om te onderzoeken in hoeverre het mogelijk was op basis van internetwaarneming representatieve cijfers over de vacaturemarkt te publiceren, zowel naar aantallen verdeeld over regio en SBI, als naar kenmerken. Achtergrond: er is behoefte aan meer gedetailleerde informatie over vacatures dan op dit moment beschikbaar is bij het CBS (Grijpstra e.a., 2010).

De internetdata, die door het UWV worden gebruikt, zijn afkomstig van het bedrijf Textkernel, die een service c.q. database Jobfeed aanbiedt. Jobfeed levert dagelijks vrijwel alle nieuwe vacatures die in Nederland op het Internet verschijnen. Daarnaast kan de gebruiker via Jobfeed Analytics inzicht krijgen in omvang en ontwikkeling van de vacaturemarkt. Bij opbouw van database Jobfeed wordt de spidering technologie van het bedrijf WiseGuys (<http://www.wiseguys.nl>) gecombineerd met Textractor van Textkernel om (volledig automatisch) alle sites in Nederland na te speuren op eventuele vacatures en deze vervolgens gestructureerd en gestandaardiseerd aan te bieden.

Het CBS heeft in de loop van 2010 – 2011 de beschikking gekregen over drie leveringen uit Jobfeed. Twee via het UWV en één rechtstreeks van Textkernel. Deze leveringen zijn door het CBS geanalyseerd.

Figuur 2. Proces van het maken van statistiek op basis van internetwaarneming (cyclisch)



4.2 Juridische aspecten

Door het toenemend gebruik van het Internet als nieuwe databron wordt het steeds duidelijker dat onderzoek naar de juridische implicaties van het verzamelen, opslaan, koppelen, bewerken en publiceren van internetdata een nog onontgonnen terrein is. De centrale vraag is of hiermee rechten van anderen worden geschonden? Zo is er de laatste jaren veel te doen over privacy en eigendomsrechten van data op het Internet.

In eerste plaats gaat het om privacy in relatie tot persoonsgegevens op het Internet. Daarbij speelt het College Bescherming Persoonsgegevens (CBP) een centrale rol. De kwalificatie “persoonsgegeven” gaat ver. Bij dit onderzoek naar het gebruik van internetvacatures gaat het specifiek om de eventuele mogelijkheid dat er éénmanszaken tussen de verzamelde internetgegevens zitten, te identificeren via onder meer KvK-nummer, bedrijfsnaam, adres, e-mailadres of IP-adres. Voor de zekerheid is er een melding gedaan bij de Functionaris Bescherming Persoonsgegevens van het CBS, het verlengstuk van het CBP (zie Heerschap, 2012A). Op verzoek van deze functionaris is ook een Privacy Impact Assessment uitgevoerd (Heerschap, 2012B). In het uiterste geval kunnen de éénmanszaken uit de verkregen bestanden worden verwijderd. Het zullen er namelijk niet veel zijn.

Dit is echter niet voldoende. Het moet ook duidelijk zijn dat het CBS het recht heeft om data van het Internet sowieso te benutten, bijvoorbeeld van intermediairs. Het gaat dan niet alleen om het verzamelen maar ook om het bewerken, koppelen en

opslaan van internetdata. Hierbij spelen bijvoorbeeld de databankenwet en het eigendomsrecht een rol. In 2009 is door de juristen van het CBS onderzoek gedaan naar de rechtmatigheid van het verzamelen, opslaan en bewerken van prijzen, die met behulp van een eigen robot op het Internet zijn waargenomen. De conclusie was dat er op zich geen beletsel was op deze wijze gegevens te verzamelen en te publiceren, mede ook omdat er in de wetgeving veel ruimte wordt gegeven aan onderzoek en statistiek (Barendswaard, 2009). De internetwaarneming voor vacatures volgt deze conclusie.

4.3 Uitgevoerde werkzaamheden

Midden 2010 heeft het CBS via het UWV de beschikking gekregen over een eerste set data van Jobfeed betreffende ontstane vacatures in het laatste kwartaal van 2009 en het eerste kwartaal van 2010 van alle bedrijfstakken, dit zonder KvK-nummers. In maart van 2011 heeft een vergelijkbare levering plaatsgevonden maar dan over het gehele jaar 2010.

Uit de analyses van deze eerste twee leveringen, met onder andere een vergelijking met de Structuurenquête van eind 2008, bleek dat het niet mogelijk was om op basis van deze data de Structuurenquête te (re)produceren of te benaderen. Extra onderzoek was gewenst en dan met name op basis van een koppeling met de Kwartaalstatistiek. Daarvoor was het nodig de KvK-nummers (bedrijfsidentificatie), die bij de vacatures horen, te verkrijgen. KvK-nummers maken het namelijk mogelijk de internetvacatures te koppelen aan de Bedrijfseenheden, de statistische eenheid van de vacaturestatistieken.

Om een bestand met internetvacatures met KvK-nummers te verkrijgen is rechtstreeks contact opgenomen met Textkernel omdat het UWV niet in staat was deze te leveren. Na lang onderhandelen en veel juridische hobbels is uiteindelijk in december 2011 een gratis derde set met internetdata ontvangen. Dit bestand was beperkt tot de eerste twee kwartalen van 2011, en dan alleen de internetvacatures in de sectoren Horeca en Financiële instellingen.

Daarbij is, gezien de resultaten van de eerste twee analyses, in de loop van het project de focus verschoven van (het produceren van) de Structuurenquête naar (produceren van) de Kwartaalstatistiek. Basis idee was om, in stappen, uiteindelijk nog maar één papieren uitvraag per jaar te doen over het aantal openstaande vacatures bij bedrijven, instellingen en overheid en de andere drie kwartalen cijfers te schatten op basis van Jobfeed-data.

Voor de resultaten van alle analyses zij verder verwezen naar paragraaf 5.

In 2012 zijn de werkzaamheden van het project overgaan van het programma Impact ICT naar een apart project Vermindering Lastendruk Vacatures en Ziekteverzuim

(VLVZ, zie Hoefkens, 2012). Zoals de naam al zegt, dit project heeft als doel om de lastendruk te verminderen. Dit project heeft een eigen stuurgroep. Het project volgt twee sporen. Bij het ene spoor wordt gekeken naar de mogelijkheid om de steekproef van de Kwartaalstatistiek te verkleinen (Daalmans, 2012), zonder gebruik te maken van aanvullende gegevens, zoals internetwaarneming. Onderzocht is welke consequenties het verkleinen van de steekproef heeft op de nauwkeurigheid van de uitkomsten. De uitkomsten van dit onderzoek hebben aanleiding gegeven om de steekproef met 5000 eenheden te verkleinen op kwartaalbasis. Bij het andere spoor wordt het traject met internetwaarneming voortgezet. Daarbij zal, evenals bij het eerste spoor, de nadruk liggen op lastendrukvermindering en veel minder op nieuwe (bijvoorbeeld meer detail) en snellere mogelijkheden van publicatie of mogelijkheden voor kwaliteitsverbetering.

5. De analyses

5.1 De eerste levering

Beschikbaar vanaf :	7 juni 2010;
Data over:	Alle bedrijfstakken;
Data periode:	Jobfeed vacatures, ontstaan in 4 ^e kwartaal van 2009 en 1 ^e kwartaal van 2010.
Inclusief KvK nummers:	Nee

In Velzen e.a. (2011) wordt uitgebreid beschreven welke analyses op dit bestand zijn uitgevoerd. Zo is onderzocht of de verdeling naar opleidingsniveau en beroep in de Jobfeed data overeenkomt met de Structuurenquête. Een moeilijkheid is dat geen recente CBS cijfers beschikbaar zijn. De meest recente data komen uit Structuurenquête 2008.

In vergelijking met de Structuurenquête 2008 is de bovenkant van de arbeidsmarkt sterk oververtegenwoordigd in de internetdata. Vacatures waarvoor een hoge opleiding wordt gevraagd komen relatief vaker voor in de Jobfeed data en hetzelfde geldt voor vacatures voor de hogere en wetenschappelijke beroepen. De verschillen tussen de CBS cijfers en de Jobfeed-cijfers zijn dusdanig groot dat het onwaarschijnlijk lijkt dat zij alleen worden veroorzaakt doordat zij op verschillende periodes zijn gebaseerd. In aanvulling op Velzen e.a., worden in tabel 2 resultaten gespecificeerd naar bedrijfstak.

Tabel 2. Procentuele verdeling naar opleidingsniveau per bedrijfstak

Branche	Structuurenquête 2008			Jobfeed 2010/2011		
	Laag	Midden	Hoog	Laag	Midden	Hoog
Landbouw en visserij	40	55	6	10	47	44
Delfstoffenwinning	13	22	65	0	40	60
Industrie	28	46	26	4	47	49
Energie- en waterleidingbedrijven	3	38	59	2	37	62
Bouwnijverheid	46	39	14	5	63	33
Handel	36	50	14	9	58	34
Horeca	58	39	3	23	55	23
Vervoer en communicatie	48	33	19	3	30	67
Financiële instellingen	14	36	50	2	36	63
Zakelijke dienstverlening	15	34	51	3	32	66
Openbaar bestuur	5	29	65	4	45	51
Onderwijs	8	22	70	2	32	66
Gezondheids- en welzijnzorg	13	54	33	10	48	42
Cultuur en overige dienstverlening	28	41	31	4	39	57

Er zijn grote verschillen tussen branches. De afwijkingen zijn relatief klein voor energie- en waterleidingsbedrijven, financiële instellingen, zakelijke dienstverlening, onderwijs en gezondheids- en welzijnzorg, terwijl deze relatief groot zijn bij de landbouw- en visserij, industrie, bouwnijverheid, handel, vervoer en communicatie, cultuur- en overige dienstverlening. Dus: voor sommige branches lijken de data van Jobfeed meer op de CBS cijfers dan voor andere branches. Opvallend is dat de branches met relatief kleine verschillen in tabel 2 meestal een relatief groot internet markt bereik in tabel 1 en vice versa hebben.

5.2 De tweede levering

Beschikbaar vanaf :	24 maart 2011;
Data over:	Alle bedrijfstakken;
Data periode:	Jobfeed vacatures, ontstaan in 2010.
Inclusief KvK nummers:	Nee

De tweede levering van Jobfeed-data omvat drie recentere kwartalen dan de eerste levering. Met behulp van dit bestand is onderzocht of het mogelijk is de Kwartaalstatistiek minder vaak uit te voeren als we de Jobfeed-gegevens zouden gebruiken.

Om deze vraag te beantwoorden is de volgende exercitie uitgevoerd:

- 1) De data van Jobfeed over het vierde kwartaal van 2009 worden opgehoogd naar het totaal van de ontstane vacatures van de Kwartaalstatistiek van het vierde kwartaal 2009.
- 2) Het ophooggewicht dat zodoende wordt verkregen (= totaal aantal vacatures in de Kwartaalstatistiek / totaal aantal Jobfeed vacatures) wordt toegepast op de

Jobfeed-cijfers van alle kwartalen van 2010. Dat wil zeggen: de Jobfeed-gegevens over de vier kwartalen van 2010 worden opgehoogd met het gewicht dat is berekend voor het vierde kwartaal van 2009.

In tabel 3 worden de schattingen, die op bovenstaande manier zijn gemaakt, vergeleken met de CBS Kwartaalstatistiek.

Tabel 3. Aantal ontstane vacatures (x 1000).

Kwartaal	Jobfeed ongewogen	Jobfeed gewogen	Kwartaalstatistiek
2009-4	107	226	226
2010-1	113	240	284
2010-2	114	241	251
2010-3	115	244	228
2010-4	129	274	254

De verschillen zijn erg groot, wat erop duidt dat de kwartaalontwikkelingen van Jobfeed niet goed overeenkomen met die van de Kwartaalstatistiek. Het is dus niet goed mogelijk om de uitkomsten van de Kwartaalstatistiek te voorspellen door extrapolatie van de uitkomsten van een eerder kwartaal met de ontwikkeling van de Jobfeed-cijfers.

Opvallend is wel dat het aantal vacatures bij Jobfeed ieder kwartaal stijgt, terwijl de Kwartaalstatistiek een daling laat zien in het tweede en derde kwartaal van 2010. Mogelijk kan dit worden verklaard door toenemend gebruik van Internet als wervingskanaal en dus een toename van het aantal postings van vacatures.

Tabel 4. Aantal ontstane vacatures (x 1000) naar SBI-1 sectie

Jobfeed – gewogen naar Kwartaalstatistiek											
		2009-4					Kwartaalstatistiek 2009-4				
SBI-1 sectie		09-4	10-1	10-2	10-3	10-4	09-4	10-1	10-2	10-3	10-4
A Landbouw, bosbouw en visserij		3	2	3	4	4	3	5	4	6	4
B Delfstoffenwinning		0	0	0	0	0	0	0	0	0	0
C Industrie		24	31	33	37	45	24	31	31	22	26
D Energievoorziening		0	0	0	0	0	0	0	1	0	0
E Waterleiding-bedrijven en afvalbeheer		1	1	1	1	1	1	1	1	1	2
F Bouwnijverheid		13	17	15	17	22	13	27	19	14	15
G Handel		52	65	65	72	84	52	56	53	50	60
H Vervoer en opslag		11	14	17	19	20	11	14	13	12	11
I Horeca		13	19	23	21	20	13	17	17	13	11

Jobfeed – gewogen naar Kwartaalstatistiek										
2009-4					Kwartaalstatistiek 2009-4					
J Informatie en communicatie	11	14	12	15	16	11	14	8	9	11
K Financiële dienstverlening	7	7	7	7	8	7	12	7	6	10
L Verhuur en handel van onroerend goed	3	3	3	3	4	3	4	3	2	2
M Specialistische zakelijke diensten	18	19	18	19	21	18	22	18	16	22
N Verhuur en overige zakelijke diensten	19	29	31	32	30	19	25	21	18	23
O Openbaar bestuur en overheidsdiensten	8	7	5	4	4	8	8	9	8	8
P Onderwijs	8	8	11	6	8	8	10	12	16	11
Q Gezondheids- en welzijnzorg	25	30	27	26	29	25	28	26	26	27
R Cultuur, sport en recreatie	4	5	4	4	5	4	4	3	3	4
S Overige dienstverlening	6	4	4	4	5	6	6	5	6	6
Totaal	226	275	279	291	326	226	284	251	228	253

De bovenstaande exercitie is nog eens herhaald, maar dan op het niveau van SBI-1 secties. Dat wil zeggen dat voor iedere SBI-1 sectie een afzonderlijke ophoogfactor is bepaald. Deze wordt weer gebaseerd op het vierde kwartaal van 2009 en toegepast op alle vier kwartalen van 2010.

De resultaten in tabel 4 laten wederom grote verschillen zien tussen de opgehoogde Jobfeed-cijfers en de CBS Kwartaalstatistiek.

Hoewel de verschillen tussen Jobfeed en de Kwartaalstatistiek over het al geheel groot zijn volgt uit tabel 4 dat de discrepantie verschilt per bedrijfstak. De verschillen zijn relatief groot bij de industrie, bouwnijverheid, handel en horeca en relatief klein bij de gezondheidszorg en cultuur sport en recreatie.

Het totaal op de onderste regel van tabel 4 wijkt af van tabel 3. Dit komt omdat vacatures met een ontbrekende SBI-code niet zijn meegenomen in tabel 4, maar wel in tabel 3. Daarnaast is op een ander aggregatieniveau opgehoogd.

De stijging van het totaal aantal Jobfeed vacatures in 2010-1 ten opzichte van 2009-4 in tabel 4 is veel groter dan in tabel 3, omdat er in het laatste kwartaal van 2009 relatief veel vacatures voorkomen zonder SBI-code. De ophoogfactoren die zijn

gebruikt in tabel 4 zijn daarom groter dan de ophoogfactoren die zijn toegepast in tabel 3.

Tabel 5. Procentuele verdeling naar beroepsniveau

	Jobfeed, geheel 2010	Structuurenquête 2008
Elementaire beroepen	2,9%	8,4%
Lagere beroepen	16,9%	26,4%
Middelbare beroepen	26,4%	39,6%
Hogere beroepen	29,1%	21,7%
Wetenschappelijke beroepen	14,1%	3,8%
Onbekend	10,6%	0,0%

De tweede dataset is ook gebruikt om de selectiviteit te onderzoeken met betrekking tot soort beroep. Tabel 5 laat zien dat de verdeling naar beroepsniveau weer sterk afwijkt van de Structuurenquête 2008. De hogere en wetenschappelijke beroepen zijn duidelijk weer oververtegenwoordigd in de internetdata.

Tabel 6. Aantal ontstane vacatures; opgehoogd Jobfeed versus CBS

Kwartaal	Jobfeed, opgehoogd	CBS
2010-1	183.166	174.000
2010-2	195.364	191.000
2010-3	200.404	197.000
2010-4	213.296	181.000

Tot slot, zijn de data van Jobfeed opgehoogd met ophoogfactoren die zijn meegeleverd in het Jobfeed bestand. Deze ophoogfactoren zijn bedoeld voor het ophogen van de Jobfeed-gegevens naar de populatie van alle Nederlandse bedrijven. De verschillen tussen de opgehoogde Jobfeed-cijfers en de Kwartaalstatistiek zijn groot, met een uitschieter voor het laatste kwartaal van 2010, zie tabel 6. Een reden voor dit grote verschil is dat de CBS cijfers over het vierde kwartaal van 2010 nog niet beschikbaar waren toen de Jobfeed dataset werd samengesteld. De ophoogfactoren zijn daarom volgens een andere, onbekende methode gemaakt. Dat het niet goed is gelukt om de cijfers uit de Kwartaalstatistiek te benaderen geeft aan dat het niet eenvoudig is om goede schattingen van het totaal aantal vacatures te maken, op basis van alleen internetwaarneming.

5.3 De derde levering

Beschikbaar vanaf :	1 december 2011;
Data over:	Horeca en Financiële instellingen Alleen die bedrijfseenheden, die tot de steekproef van de Kwartaalstatistiek behoren
Data periode:	Jobfeed vacatures, ontstaan in het 1 ^e en 2 ^e kwartaal 2011.

Inclusief KvK nummers: Ja

Het bijzondere aan deze derde levering is dat voor de eerste keer Jobfeed data zijn aangeleverd, verrijkt met Kamer van Koophandelnummers. Dit maakte het voor het eerst mogelijk om de Jobfeed-gegevens op bedrijfsniveau te koppelen aan de Kwartaalstatistiek.

Omdat verwacht werd dat de kosten van een uitgebreide levering (te) hoog zouden zijn, is er voor gekozen om alleen de Jobfeed data van twee branches op te vragen: de horeca en de financiële instellingen. Voor deze twee branches is gekozen, omdat verwacht werd dat deze erg verschillend zijn, voor wat betreft de vergelijkbaarheid met de Kwartaalstatistiek. Het vermoeden was dat de Jobfeed data van de financiële instellingen relatief veel zouden lijken op de Kwartaalstatistiek, terwijl de verschillen juist groot zouden zijn bij de horeca. Dit vermoeden is gestaafd op tabel 2, die laat zien dat voor de financiële instellingen de verdeling naar opleidingsniveau relatief veel op de verdeling in de Structuurenquête 2008 lijkt, terwijl de verschillen bij de horeca juist behoorlijk groot zijn.

Het koppelingsproces is als volgt verlopen. Eerst zijn alle bedrijfsindentificatienummers (beids), die aan het einde van het tweede kwartaal, in de respons van de Kwartaalstatistiek voorkwamen in de branches Horeca en Financiële instellingen geselecteerd. Het gaat om 1159 beids bij de horeca en 957 beids bij de financiële instellingen. Daarna zijn door het ABR de daarbij behorende KvK-nummers opgezocht. Bij de horeca zijn er 9 beids waaraan geen KvK-nummers konden worden gekoppeld. Bij de financiële instellingen zijn aan alle 957 beids KvK-nummers gekoppeld. Het CBS-bestand met KvK-nummers is aan Textkernel verstrekt. Textkernel heeft vervolgens alle vacatures uit de Jobfeed database van de eerste twee kwartalen van 2011, die koppelde aan de verstrekte KvK-nummers, terug geleverd aan het CBS³. Ten slotte zijn de Jobfeed-gegevens door het CBS weer geaggregeerd van KvK-nummers naar beids. Op brancheniveau komen we dan tot de volgende resultaten

Tabel 7. Aantallen vacatures (JF=Jobfeed, onopgehoogd)

Branche	Soort	Kwartaal 1			Kwartaal 2		
		Ruw*	Gaaf*	JF	Ruw*	Gaaf*	JF
Fin.inst.	Ontstaan	1467	1683	2170	2189	2280	2259
Horeca	Ontstaan	1966	2370	1689	9447	3173	2086
Fin.inst.	Verdwenen	1451	1253	1160	2074	1976	1710
Horeca	Verdwenen	2244	2199	631	3362	3042	1452
Fin.inst.	Open	3413	3413	1010	4657	4657	1559
Horeca	Open	1661	1546	1058	2182	2172	1692

**CBS resultaten; JF staat voor Jobfeed.*

³ Textkernel koppelt KvK-nummers op basis van naam, adres e.d. aan gevonden vacatures met behulp van specifiek aangekochte software en verkregen gegevens van de Kamer van Koophandel.

Tabel 7 laat zien dat

- de verschillen tussen Jobfeed en de Kwartaalstatistiek groter zijn bij de horeca dan bij de financiële instellingen (zoals verwacht);
- het aantal Jobfeed vacatures meestal lager is dan het aantal vacatures volgens het CBS. Dit is niet vreemd; omdat Jobfeed in principe alleen de vacatures waarneemt, die op internet worden gepubliceerd, terwijl de doelpopulatie van de Kwartaalstatistiek alle vacatures is.
- de verschillen relatief klein zijn bij de ontstane vacatures van de financiële instellingen. Dit geldt met name voor kwartaal 2.
- de veranderingen in het aantal vacatures tussen kwartaal 1 en 2 volgens Jobfeed duidelijk afwijkt van het CBS.
- het aantal openstaande vacatures volgens Jobfeed veel kleiner is dan volgens het CBS. Eén van de redenen is dat het CBS alleen Jobfeed gegevens heeft ontvangen over vacatures die zijn ontstaan vanaf het eerste kwartaal van 2011. We missen dus de gegevens over vacatures die vóór 1-1-2011 zijn ontstaan en nog steeds openstaan. Het is dus van belang om data van een langere periode te hebben om zinnige uitspraken te kunnen doen over aantallen openstaande vacatures.

In tabel 8 worden de vacaturegegevens vergeleken op bedrijfsniveau. Er zijn 2.031 bedrijfseenheden gekoppeld. Tabel 8 laat zien dat het aantal vacatures per bedrijfseenheid volgens de Kwartaalstatistiek meestal groter is dan het aantal volgens Jobfeed. In de Kwartaalstatistiek zijn er ook minder bedrijfseenheden waar geen vacatures zijn ontstaan in de twee onderzochte kwartalen.

Hoewel de uitkomsten van het CBS en Jobfeed op enkele onderwerpen fors verschillen (zie tabel 7), zien we overeenkomsten bij bedrijfseenheden waar geen vacatures zijn ontstaan. Er zijn relatief veel bedrijfseenheden, die op grond van beide bronnen, Jobfeed en het CBS, geen vacatures hebben.

Op grond van de Jobfeed data kan redelijk goed voorspeld worden welke bedrijfseenheden geen vacatures hebben in de Kwartaalstatistiek. Er zijn namelijk 1746 bedrijfseenheden zonder Jobfeed vacatures en een groot deel hiervan, 1427 beid's, hebben ook geen vacatures volgens het CBS.

Tabel 8. Waarnemingen ontstane vacatures. (gesommeerd over 2 kwartalen)

Waarneming	Aantal
Aantal Beids	2031
CBS = Jobfeed	1453
w.v. CBS = Jobfeed = 0	1427
w.v. CBS = Jobfeed > 0	26
CBS ≠ Jobfeed	578
w.v. CBS > Jobfeed	426
w.v. CBS < Jobfeed	152
Jobfeed =0	1746
w.v. CBS =0	1427
w.v. CBS >0	319
CBS =0	1508
w.v Jobfeed =0	1427
w.v Jobfeed > 0	81

Bij de bedrijfseenheden met minimaal één vacature zijn de verschillen tussen Jobfeed en de Kwartaalstatistiek juist groot. Van alle bedrijfseenheden met minimaal 5 vacatures zijn er in kwartaal 1 slechts 11 waarvoor het verschil tussen Jobfeed en de Kwartaalstatistiek kleiner is dan 30% en in kwartaal 2 zijn dit er slechts 10.

We focussen nu op het deel van de CBS steekproef dat volgens Jobfeed minimaal één ontstane vacature heeft. In tabel 9 staat in de kolom “CBS selectie” het totaal aantal vacatures volgens de Kwartaalstatistiek Vacatures van deze bedrijfseenheden. Dit aantal wordt vergeleken met het totaal aantal vacatures van alle bedrijfseenheden, dat wordt weergegeven in de kolom “CBS alles”. De absolute en relatieve verschillen tussen “CBS selectie” en “CBS alles” staan in de kolommen “Fout abs” en “Fout %”.

Opvallend is dat bij de ontstane en verdwenen vacatures de procentuele fout in kwartaal 2 meestal niet veel afwijkt van kwartaal 1. Dus, de kwartaalontwikkeling, van alle bedrijfseenheden met minimaal één Jobfeed vacature, lijkt veel op de kwartaalontwikkeling van alle bedrijfseenheden. Dit komt ook naar voren uit tabel 10. In deze tabel wordt een schatting gemaakt voor de vacaturetotalen van het tweede kwartaal, door het resultaat van het eerste kwartaal te vermenigvuldigen met de ontwikkelingen van alle bedrijfseenheden met minimaal één Jobfeed vacature. Bijvoorbeeld: In kwartaal 1 zijn er in totaal 1683 vacatures ontstaan bij de financiële instellingen (zie tabel 7). De kwartaalontwikkeling van “CBS selectie”: de groep bedrijfseenheden met minimaal 1 Jobfeed vacature, is (2142 / 1541). Het totaal aantal openstaande vacatures in kwartaal 2 wordt dan geschat als: $1683 * (2142 / 1541) = 2339$. Ter vergelijking: in de kolom “CBS gaaf” staat het vacaturetotaal van de gehele populatie: 2280.

Tabel 9. Vacaturetotalen; gebaseerd op bedrijfseenheden met ≥ 1 Jobfeed vacature

Soort		Kwartaal 1				Kwartaal 2			
		CBS,	CBS	Fout	Fout	CBS,	CBS	Fout	Fout
		selectie	Alles	abs	%	selectie	Alles	Abs	%
(JF ≥ 1)									
F	Ontstaan	1541	1683	-142	8	2142	2280	-138	6
H	Ontstaan	1702	2370	-668	28	2299	3173	-874	28
F	Verdwenen	1097	1253	-156	12	1831	1976	-145	7
H	Verdwenen	1442	2199	-757	34	1789	3042	-1253	41
F	Open	3301	3413	-112	3	4500	4657	-157	3
H	Open	1081	1546	-465	30	1786	2172	-386	18

Tabel 10. Schattingen en realisaties voor kwartaal 2

Branche	Soort	Schatting	CBS Gaaf	Abs Fout	Rel.Fout(%)
Fin.inst.	Ontstaan	2339	2280	59	3
Horeca	Ontstaan	3201	3173	28	1
Fin.inst.	Verdwenen	2091	1976	115	6
Horeca	Verdwenen	2728	3042	-314	10
Fin.inst.	Open	4653	4657	-4	0
Horeca	Open	2554	2172	382	18

De schattingsfouten in tabel 10 zijn relatief klein: dus de ontwikkeling van alle bedrijfseenheden kan goed worden geschat op basis van de bedrijfseenheden met minimaal één Jobfeed vacature (met uitzondering van de verdwenen en openstaande vacatures bij de horeca).

In tabel 11 is te zien dat verreweg de meeste bedrijfseenheden geen vacatures hebben; het aantal bedrijfseenheden met minimaal 1 vacature is dus relatief klein.

Tabel 11. Aantallen bedrijfseenheden in de Kwartaalstatistiek

Branche	Kwartaal 1		Kwartaal 2	
	Totaal	≥ 1 JF vacat.	Totaal	≥ 1 JF vacat.
Fin.inst.	916	46	957	52
Horeca	1115	168	1147	235

De conclusie van het bovenstaande is dat fors op het aantal vragenlijsten kan worden bespaard wanneer gekozen wordt voor de strategie, waarin:

- Eén kwartaal per jaar de volledige CBS steekproef wordt uitgezonden;
- Drie kwartalen per jaar alleen vragenlijsten worden uitgezonden aan de bedrijfseenheden, met minimaal 1 Jobfeed vacature.

De kwartaalontwikkelingen van de groep bedrijfseenheden met minimaal één Jobfeed vacature wordt dan gebruikt voor het schatten van de vacaturetotalen voor de kwartalen waarvoor geen volledige steekproef wordt uitgezonden. Stel bijvoorbeeld

dat in kwartaal 1 de volledige steekproef wordt uitgezonden. De vacaturetotalen van kwartaal 2 wordt dan geschat door het totaal aantal vacatures in kwartaal 1 te vermenigvuldigen met de ontwikkeling van de aantallen vacatures van de bedrijfs-eenheden met minimaal 1 Jobfeed vacature.

Bij het bovenstaande valt echter wel de kanttekening te maken dat de conclusie is getrokken op basis van een vergelijking van slechts twee kwartalen en dus één kwartaalverandering. Om een betrouwbaarder beeld te krijgen is het aan te bevelen de bovenstaande exercitie te herhalen op data die meer kwartalen omvatten.

Tot slot is ook voor de derde dataset onderzocht wat de verdeling is van het aantal vacatures naar opleidingsniveau. De resultaten in tabel 12 geven hetzelfde beeld als hiervoor, namelijk dat de bovenkant van de arbeidsmarkt is oververtegenwoordigd in de internet waarneming. Ter vergelijking: de percentages laag, middel en hoog opleidingsniveau in de structuurenquête 2008 zijn: 14%, 36%, 40% voor de financiële instellingen en 58%, 39% en 3% voor de horeca (zie tabel 2).

Tabel 12. Ontstane vacatures naar opleidingsniveau; Jobfeed

Branche	Kwartaal 1			Kwartaal 2		
	Laag	Middel	Hoog	Laag	Midden	Hoog
Fin.inst.	2	26	72	1	23	77
Horeca	32	53	15	30	55	15

6. Conclusies en vervolg

6.1 Conclusies

Het CBS heeft in de periode 2010 – 2011 drie bestanden met Jobfeed data geanalyseerd. Doelstellingen waren: 1) om de papierenwaarneming van de Structuurenquête te vervangen door Jobfeed data en 2) om de uitvraag bij de Kwartaalstatistiek voor een aantal kwartalen te baseren op Jobfeed data.

De bevindingen zijn:

- In vergelijking met de Structuurenquête 2008 zijn vacatures aan de bovenkant van de arbeidsmarkt oververtegenwoordigd in de Jobfeed data. Er zijn echter grote verschillen tussen bedrijfstakken: voor sommige bedrijfstakken zijn de verschillen duidelijk groter dan voor andere.
- De ontwikkelingen en de niveaus van de vacaturetotalen van Jobfeed op macro-niveau sluiten niet goed aan bij de Kwartaalstatistiek. Dit betekent dat de uitkomsten van de reguliere Kwartaalstatistiek niet goed kunnen worden “voorspeld” door extrapolatie van de uitkomsten van een eerder kwartaal met de kwartaalverandering van de Jobfeed-gegevens.

Er was wel een uitzondering op het bovenstaande. Voor één kwartaal en één branche (financiële instellingen) zijn de verschillen tussen het totaal aantal ontstane vacatures van Jobfeed en het CBS opvallend klein. Onduidelijk is of dit door toeval komt of niet. Meer Jobfeed-gegevens zijn noodzakelijk om hier uitsluitsel over te geven.

- Als het gaat om het aantal vacatures op microniveau zijn de verschillen tussen de internetwaarneming en de CBS waarneming meestal aanzienlijk.
- Hoewel de uitkomsten van het CBS en Jobfeed op enkele onderwerpen fors verschillen, zien we overeenkomsten bij bedrijfseenheden waar geen vacatures zijn ontstaan. Bedrijfseenheden waar geen vacatures zijn ontstaan volgens de Kwartaalstatistiek, hebben meestal ook geen ontstane vacatures volgens Jobfeed, en vice versa. Op grond van de Jobfeed data kan dus redelijk goed voorspeld worden welke bedrijfseenheden geen vacatures hebben in de Kwartaalstatistiek.
- De kwartaalontwikkelingen uit de Kwartaalstatistiek worden goed benaderd door de kwartaalontwikkelingen van de groep van bedrijfseenheden die op grond van Jobfeed één of meerdere vacatures hebben. Een kansrijke strategie is dan ook om één van de vier kwartalen de volledige Kwartaalstatistiek uit te sturen en om in de drie overige kwartalen alleen vragenlijsten uit te sturen aan de bedrijfseenheden die volgens Jobfeed minimaal één vacature hebben. De ontwikkeling van het aantal vacatures van de bedrijven met minimaal één Jobfeed vacature wordt dan gebruikt als proxy voor de ontwikkeling van het totaal aantal vacatures. Bovenstaande strategie zou een enorme vermindering van de enquêtedruk betekenen, aangezien slechts een klein deel van alle bedrijfseenheden één of meerdere Jobfeed vacatures heeft. De strategie zou echter tot vertekende uitkomsten kunnen leiden en vereist daarom nader onderzoek. Een tweede reden voor verder onderzoek is dat bovenstaande conclusie is gebaseerd op de data van slechts twee kwartalen. Het is daarom goed denkbaar de bevinding gebaseerd is op toeval.
- Vastgesteld moet worden dat de waarneming van vacatures via internetrobots en de structurering daarvan (nog) hiaten bevatten die op dit moment moeilijk te kwantificeren zijn. Te noemen zijn de sterke toename van het gebruik van het Internet als wervingskanaal voor personeel (dus toename van het aantal postings), (de periode van) het ontdubbelen, vacatures van intermediairen en uitzendbureaus en bulkvacatures. Textkernel is voortdurend bezig om het proces te verbeteren. Daarbij moet worden aangetekend dat de waarneming via papierenvragenlijsten ook niet zonder problemen is.
- Dat er zich vooralsnog geen juridische problemen voordoen als het gaat om het verkrijgen, opslaan, bewerken en koppelen van internetdata over vacatures. Als het gaat om privacy gevoelige gegevens, zoals die van

éénmanszaken, dan kunnen die in het uiterste geval worden verwijderd uit de datasets.

Hiermee moet worden geconstateerd dat de vooraf en in de loop van het project gestelde doelstellingen vooralsnog niet haalbaar zijn, althans – voor de duidelijkheid - met de resultaten van de nu uitgevoerde analyses. Wel zijn er aanknopingspunten voor vervolgonderzoek.

De beste optie lijkt op dit moment de optie waarbij geen vraaglijsten worden uitgestuurd naar bedrijven die in de Jobfeed data geen vacatures hebben. Daarbij moet wel worden onderzocht of dit qua tijdschema past in de eis dat 45 dagen na afloop van het kwartaal de cijfers beschikbaar moeten zijn. Een aanpak zou kunnen zijn om in een vroegtijdig stadium al vragenlijsten uit te sturen naar die bedrijven die op dat moment al wel vacatures bij Jobfeed hebben.

6.2 Vervolg

De bovenstaande bevindingen zijn gebaseerd op de analyse van drie datasets. Het betrof verkennend onderzoek. De hoeveelheid informatie in deze drie datasets is relatief beperkt geweest. Slechts voor twee kwartalen en twee bedrijfstakken was het mogelijk Jobfeed-gegevens op microniveau aan de Kwartaalstatistiek te koppelen. Om meer gedegen uitspraken te kunnen doen over de bruikbaarheid van de Jobfeed-gegevens is het nodig om op grotere schaal Jobfeed data te koppelen aan de Kwartaalstatistiek. Dit is onder meer van belang voor het bepalen van het aantal openstaande vacatures op één tijdstip. Hiervoor is het nodig om informatie te hebben over vacatures die eerder zijn ontstaan en nog steeds openstaan.

Voor verder onderzoek wordt gedacht aan:

- Aanvulling op al uitgevoerde analyses:
 - a) Onderzoeken of de verschillen tussen de internetwaarneming en de Kwartaalstatistiek afhangen van de grootteklasse. Lijkt de internetwaarneming van de grotere bedrijven meer op de CBS waarneming, of andersom bij het berekenen van de macrototalen, de gegevens per grootteklasse op te hogen, zoals in de praktijk gebeurt. In de analyses die tot nu toe zijn uitgevoerd is dit niet gebeurd.
 - b) Uitwerken van de optie om geen enquêteformulieren uit te sturen als er bij Jobfeed sprake is van 0 vacatures. Dit kan de administratieve lasten al aanzienlijk verminderen. Het proces moet wel voldoen aan de eis dat gegevens 45 dagen na afloop van het kwartaal beschikbaar zijn.
- Tot nu toe zijn de uitgevoerde analyses vooral gericht op het niveau van Nederland totaal en SBI-1 digit. Dit kan worden uitgebreid met vergelijkbare analyses op microniveau.

- Het benutten van extra informatie. Zo komen er in 2012 meer recente gegevens van het UWV beschikbaar over het percentage vacatures dat bedrijven op het Internet zetten. Mogelijk zijn er ook nog andere bestanden die extra informatie leveren. Er kan ook gedacht worden om bij één van de uitzendingen van de Kwartaalstatistiek meer (hulp)informatie te vragen, die gebruikt kan worden bij samenstellen van de andere drie Kwartaalstatistieken op basis van Jobfeed data.
- Door te kijken of er opties zijn waarbij het gebruik van Jobfeed data niet zo zeer leidt tot het totaal vervangen van de uitvraag maar eerder tot een substantiële beperking van de steekproef.

Voorgesteld wordt om te beginnen met:

1. Het aanschaffen van twee jaar Jobfeed data, inclusief de koppeling aan KvK-nummers.
2. Het organiseren van een brainstormsessie (1 hele dag) met experts op het terrein van vacatures, vacatures op het internet, methodologie, spidering en ontdebelling van vacatures en het kader (ABR). Deelname hoeft niet beperkt te blijven tot CBS'ers. Doel is het uitwerken van enkele kansrijke scenario's voor verder onderzoek⁴.
3. Concrete uitwerking en uitvoering van de meest kansrijke scenario's.

Ten slotte zij nog opgemerkt dat de keuze voor vervolgonderzoek bepaald wordt door de doelen die worden nagestreefd. Is het doel alleen lastendrukvermindering? Wil men ook kijken naar snellere cijfers of meer gedetailleerd, vooral aansluitend bij de behoeften in de buitenwereld (zie o.a. Grijpstra e.a., 2010). Of "beperken" we ons tot de informatiebehoefte van Eurostat? En, ten slotte. is kwaliteitsverbetering een issue?

⁴ Het EIM biedt goede faciliteiten om dit te organiseren.

7. Referenties

- Barendswaard, P. (2009), Internetrobots en wetgeving (Databankenwet, CBS-wet), 2-10-2009.
- Binnenlands Bestuur (2012), Nieuwe portal maakt CBS cijfers makkelijker vindbaar, 20-01-2012.
- Daalmans, J. (2012), Steekproefadvies 2012 vacature- en ziekteverzuimstatistiek, interne nota CBS
- Drees, M. (2012A), Jobboard landschap verandert snel, kansen voor recruiters, column op het Internet, 23 januari 2012.
- Drees, M. (2012B), De opkomst en ondergang van Monsterboard, column op het Internet, 14 januari 2012.
- Dialogic (2008), Go with the data flow, 13 mei 2008.
- Grijpstra, D.H., G. Mulder en M. Folkeringa (2010), Eindverslag pilot Vacaturemeter, september 2010.
- Heerschap N. (2010A), Inzet van internetrobots bij de Vacature-enquête, Project Initiatie Document (klein), Versie 1.1
- Heerschap, N. (2010B), Programma Impact ICT 2011, september 2010.
- Heerschap, N. (2011), Programma Impact ICT 2012, juni 2011.
- Heerschap, N. (2012A), Melding internetrobots.
- Heerschap, N. (2012B), Privacy Impact Assessment internetrobots.
- Hoefkens R., (2012), Vermindering Lastendruk vacatures en ziekteverzuim, Project Initiatie document, versie 1.0p2
- Hoekstra, R., O. ten Bosch en F. Harteveld (2010), Automated data collection from web sources for official statistics: first experience, 25 april 2010.
- Intermediair (2011), Vacaturesites de belangrijkste bron voor werkzoekenden, 30-08-2011
- Ruth Van, F.J. en R. Wekker (2009), Monthly employers' sentiment indicators; doing more with business survey data, Centraal Bureau voor de Statistiek discussionpaper 09009. Url: <http://www.cbs.nl/NR/rdonlyres/1459ED14-E741-4384-94B4-1252AA4B11F7/0/200909x10pub.pdf>
- Ruth Van, F.J. en R. Wekker (2011), Nieuwe Vacatureindicatoren; Doel, Achtergrond en Samenstellingen, Rapport, Centraal Bureau voor de Statistiek. Url: <http://www.cbs.nl/NR/rdonlyres/3A3DB463-D4E1-4BC2-A279-F3189C402DF5/0/2011nieuwevacatureindicatoren.pdf>

Velzen M. van, J.A. Daalmans, N. Heerschap, D. ter Steege (2011), Statistieken van de vacatures op internet: een vooronderzoek, Interne nota DMV-2011-07-28-MVLN-JDAS-NHRP-DSGE CBS Den Haag.

BIJLAGE I. Achtergrond Jobfeed database

Textkernel behoort tot de bedrijven die het Internet afzoeken naar vacatures en deze gestructureerd en ontdebeld in een database, in dit geval Jobfeed, onderbrengen en deze vacature-informatie aan derden ter beschikking stellen voor bijvoorbeeld marktonderzoek of statistiek. Bij Textkernel gaat het spideren in samenwerking met het bedrijf Wiseguys. Daarbij gaat het om:

- Websites die specifiek gespiderd worden. Het betreft hier met name jobboards, intermediairen en kranten.
- Websites die wild worden gespiderd. Dit is een doorlopend proces. Daarbij wordt in het Nederlandstalige deel van het Internet gezocht op tekstkenmerken.

De stappen die worden gevolgd om Jobfeed te vullen, zijn:

- 1 *Spideren*, waarbij het Internet systematisch wordt afgezocht naar informatie over vacatures.
- 2 *Classificatie*. Hierbij wordt vastgesteld of een opgehaalde internetpagina een vacature is. Bij de specifiek gespiderde sites geldt dat natuurlijk niet omdat daar per definitie sprake is van vacatures. In dit stadium wordt ook al gekeken naar dubbele vacatures, namelijk als sprake is van dezelfde url van een al eerder waargenomen vacature. In dat geval wordt de vacature niet meer meegenomen.
- 3 *Informatie-extractie*. Met behulp van een module (Texttractor) wordt:
 - a. De inhoud van de webpagina ontdaan van niet relevante inhoud. Hierdoor blijft alleen de werkelijke vacaturetekst over.
 - b. Uit de overgebleven tekst wordt de relevante informatie geëxtraheerd en gevalideerd voor variabelen zoals vacaturetitel, standplaats, opleiding en organisatie.
- 4 *Normalisatie en matching*. In deze stap wordt de tekst van de variabelen gestandaardiseerd op basis van standaardindelingen. Dit gebeurt onder andere voor plaats, beroep, opleiding en organisatie. Wat de organisatie betreft wordt op basis van gegevens als plaats, naam en adres gekoppeld aan het Kamer van Koophandelnummer. Hierdoor is het mogelijk om gegevens, zoals SBI en grootteklasse toe te voegen aan de informatie over de vacature.
- 5 *Ontdubbeling*. Daarbij wordt onderscheid gemaakt tussen twee typen dubbelen:

- a. Dubbele vacatures die op dezelfde site zijn gevonden, bijvoorbeeld die dubbel geplaatst zijn of die bereikbaar zijn via verschillende url's;
- b. Dubbele vacatures, die gevonden zijn op verschillende sites. Bijvoorbeeld geplaatst bij de website van een bedrijf en op een jobboard.

Bij het ontdebellen, het vergelijken van vacatures, gaat men drie weken terug. Op dit moment is men aan experimenteren met zes weken. Het ontdebelleningsproces verloopt via een vaste set met regels (business rules). Met de melding dat een vacature verlopen is wordt niets gedaan. Pas als de vacature niet meer wordt aangetroffen wordt hij als verlopen beschouwd. Als dezelfde vacature later weer op dezelfde site verschijnt blijft de datum verlopen staan. Betreft het een andere site dan wordt het beschouwd als een nieuwe vacature.

- 4. *Beschikbaar stellen van de database Jobfeed.*
- 5. *Monitoring, kwaliteitscontroles en verbetering.*