

Waarnemingsinnovatie: Nieuwe bronnen en mogelijkheden



Marko R. Roos, Piet J.H. Daas en Marco Puts

De in dit rapport weergegeven opvattingen zijn die van de auteurs en komen niet noodzakelijk overeen met het beleid van het Centraal Bureau voor de Statistiek

Discussion paper (09027)



Verklaring van tekens

.	= gegevens ontbreken
*	= voorlopig cijfer
x	= geheim
—	= nihil
—	= (indien voorkomend tussen twee getallen) tot en met
0 (0,0)	= het getal is kleiner dan de helft van de gekozen eenheid
niets (blank)	= een cijfer kan op logische gronden niet voorkomen
2007–2008	= 2007 tot en met 2008
2007/2008	= het gemiddelde over de jaren 2007 tot en met 2008
2007/'08	= oogstjaar, boekjaar, schooljaar enz., beginnend in 2007 en eindigend in 2008
2005/'06–2007/'08	= oogstjaar, boekjaar enz., 2005/'06 tot en met 2007/'08

In geval van afronding kan het voorkomen dat het weergegeven totaal niet overeenstemt met de som van de getallen.

Colofon

Uitgever

Centraal Bureau voor de Statistiek
Henri Faasdreef 312
2492 JP Den Haag

Prepress

Centraal Bureau voor de Statistiek - Graimedia

Omslag

TelDesign, Rotterdam

Inlichtingen

Tel. (088) 570 70 70
Fax (070) 337 59 94
Via contactformulier: www.cbs.nl/infoservice

Bestellingen

E-mail: verkoop@cbs.nl
Fax (045) 570 62 68

Internet

www.cbs.nl

ISSN: 1572-0314

© Centraal Bureau voor de Statistiek, Voorburg/Heerlen, 2009.
Verveelvoudiging is toegestaan, mits het CBS als bron wordt vermeld.

Waarnemingsinnovatie: nieuwe bronnen en mogelijkheden

Marko R. Roos, Piet J.H. Daas en Marco Puts

Centraal Bureau voor de Statistiek, Divisie Methodologie en Kwaliteit,
Postbus 4481, 6401 CZ Heerlen; mros@cbs.nl, pdas@cbs.nl, mjps@cbs.nl.

Samenvatting: Dit rapport geeft een overzicht van nieuwe secundaire databronnen die voor het CBS mogelijk interessante gegevens bevatten. Voor een aantal van die bronnen is een statistische toepassing denkbaar. Ook worden een aantal nieuwe toepassingen besproken. Het belangrijkste doel van dit rapport is een inspirerende en verfrissende kijk te geven op het gebruik van secundaire bronnen door het CBS.

Trefwoorden: Secundaire bronnen, XBRL, Geografische gegevens, GSM, Internet, databronnen

1. Inleiding

Het Centraal Bureau voor de Statistiek (CBS) wil onbetwiste statistieken over de Nederlandse samenleving op een doelmatige en doelgerichte wijze samenstellen. Om dit doel te bereiken wil zij zoveel mogelijk gebruik maken van administratieve gegevens die in Nederland beschikbaar zijn. Dit betreft zowel gegevens die binnen de overheid als gegevens die binnen Nederlandse organisaties aanwezig zijn (CBS 2004). Uit kostenoverweging en vanuit het oogpunt van de administratieve lastendruk, is het belangrijk dat het CBS probeert zo veel mogelijk informatie uit administratieve bronnen te gebruiken. Dit geldt niet alleen voor al bestaande statistieken maar ook voor eventuele nieuwe statistische publicaties. Gezien het belang van het verlagen van de administratieve lastendruk door de overheid valt te

verwachten dat het publiceren van nieuwe statistieken door het CBS in de nabije toekomst vrijwel uitsluitend mogelijk is door:

- i. reeds door het CBS verzamelde informatie ‘slim’ te combineren
- ii. kosteloos of zo goedkoop mogelijk gebruik te maken van reeds door anderen verzamelde gegevens (secundaire bronnen)
- iii. gebruik te maken van publiek beschikbare gegevens.

Deze constatering vormen het uitgangspunt voor het programma ‘Waarnemingsinnovatie: nieuwe bronnen’ van Methodologie. De laatste twee punten worden in dit rapport verder uitgewerkt. Het ‘slim’ hergebruiken van reeds verzamelde informatie behoort nadrukkelijk niet tot de scope van het (in dit rapport beschreven) werk voor een programma ‘Waarnemingsinnovatie: nieuwe bronnen’.

Het CBS maakt momenteel al veel gebruik van gegevens uit secundaire bronnen, maar hier is nog een aanzienlijke verbetering mogelijk. Dit rapport wil een verfrissende kijk geven op het gebruik van nieuwe secundaire bronnen door het CBS en als inspiratie dienen voor de statistische medewerkers van het CBS.. Het dient als startpunt voor het op te zetten programma ‘Waarnemingsinnovatie: nieuwe bronnen’. Dit programma levert voor het CBS nieuwe ideeën en praktische ervaringen op over de statistische (on)mogelijkheden van data die verzameld wordt door derden. Dergelijke gegevens dienen beschrijvend te zijn voor een veelheid van de economische en maatschappelijke activiteiten in Nederland. Net als voor alle andere gegevens die het CBS verzameld, is het hierbij belangrijk de privacy aspecten van personen en bedrijven te waarborgen (CBS 2004).

Het belangrijkste doel van dit rapport is het inventariseren van de mogelijkheden van nog niet door het CBS gebruikte databronnen. Het zou prettig zijn wanneer voor dergelijke bronnen direct een bestaande statistische toepassing te bedenken valt. In de tabel in Bijlage 1 is daartoe een poging gedaan. In deze tabel is geprobeerd de nieuwe databronnen te relateren aan het mogelijke gebruik voor publicaties, ingedeeld volgens de CBS thema-indeling. Omdat in dit rapport veelvuldig van afkortingen gebruikt wordt gemaakt is in Bijlage 2 een lijst met afkortingen en betekenissen opgenomen.

2. Waarnemingsinnovatie

In onze huidige samenleving worden van veel van onze dagelijkse activiteiten gegevens opgeslagen. Wat gebeurt er bijvoorbeeld wanneer je op een zaterdagochtend naar de supermarkt loopt en daar boodschappen doet. Ter eerste is de kans aanwezig dat (een deel van) je wandeling door camera’s wordt vastgelegd. In grote steden is dit eerder regel dan uitzondering (Homburg en Dekkers 2003, Koning 2008). Dit kan overigens ook de ‘webcam’ van je buurman zijn. Ook in de supermarkt wordt je bezoek met videocamera’s vastgelegd. Het telefoontje dat je tijdens de wandeling pleegde laat eveneens sporen na. Hierdoor is dan weer redelijk exact te bepalen waar je je op het moment van bellen bevond. Preciezer, wanneer je

mobiele telefoon gedurende de hele wandeling aanstond (en bij wie is dat tegenwoordig niet het geval?) is de gehele route te achterhalen (zie paragraaf 2.1.4). Bij het afrekenen van de boodschappen worden ook gegevens bewaard. Zo worden de prijs en hoeveelheid van de gekochte producten opgeslagen maar ook het tijdstip en het totaal gespendeerde bedrag. Wanneer je elektronisch betaalt ('pinnen') worden de locatie, het tijdstip en het totaalbedrag ook bij de bank opgeslagen (Betaalwijzer 2006). Bij houders van een klantenkaart worden naast de klantgegevens en de spaarpunten (bijvoorbeeld airmiles) ook de product- en prijsinformatie opgeslagen. Dit eenvoudige voorbeeld maakt duidelijk dat het niet verwonderlijk is dat burgers (en bedrijven) soms door een 'Big Brother is watching you' gevoel overvallen worden.

Wanneer er verantwoord met de privacygevoeligheid van de verzamelde informatie wordt omgegaan (Koning 2008, Mijnprivacy 2008) zitten er ook veel voordelen aan de opslag van de gegevens. Zo kan de schuldvraag van het ongeluk dat die ochtend plaatsvond door het cameratoezicht eerlijk worden beantwoord, kan de mobiele telefoonmaatschappij -door je belgedrag te monitoren- je een goedkoper abonnement aanraden en krijg je korting op je aankopen door een klantenkaart te gebruiken. Bedrijven en instellingen, personen en huishoudens laten bij hun gedrag steeds vaker een digitale voetafdruk achter. Voor een instituut als het CBS dat graag wil publiceren over de economische en sociale activiteiten van de Nederlandse samenleving is de alom aanwezige data een potentieel zeer interessante bron van informatie. De uitdaging voor een instituut als het CBS is om, vanuit de zorgvuldigheid waarmee altijd gewerkt wordt om de privacy van personen en bedrijven zo goed mogelijk te waarborgen (CBS 2004), meer gebruik te maken van digitaal beschikbare gegevens. Onderzoek hiernaar zal dan ook naast het bepalen van de bruikbaarheid van de gegevens, focussen op het waarborgen van de privacy van de Nederlandse burgers en bedrijven.

2.1 Potentiële databronnen

In het hiervoor besproken voorbeeld werden al een aantal informatiebronnen genoemd. Er zijn er echter veel meer. Van de momenteel beschikbare bronnen zijn er voor het CBS een aantal als potentieel interessant aan te wijzen:

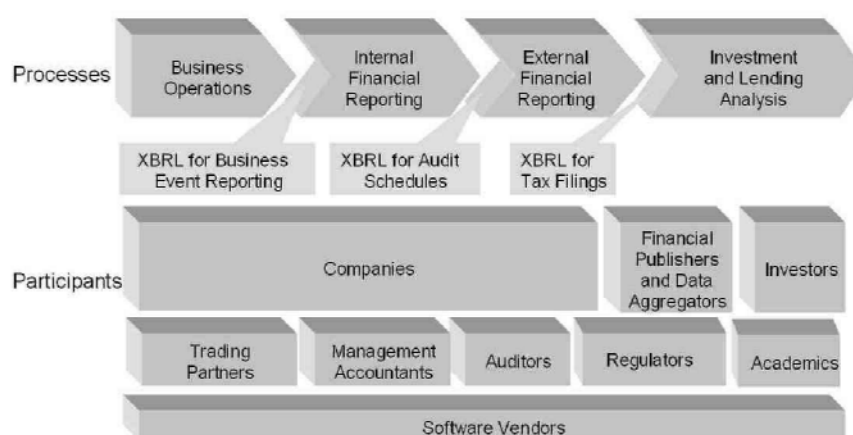
- Bedrijfsadministraties
- Radio frequency identification tags (RFID's)
- Satelliet gebaseerde navigatiesystemen (Global positioning systems; GPS)
- Mobiele telefonie gegevens (telefoonmastgegevens; GSM)
- Internet.

De hierboven vermelde databronnen en hun gebruik worden in de volgende paragrafen verder besproken. Waar mogelijk zullen ook één of meer mogelijke toepassingen van die gegevens door het CBS worden genoemd. Aan het eind van dit hoofdstuk worden nog een aantal toekomstige mogelijk interessante databronnen

genoemd. Ten slotte is in de tabel in bijlage 1 geprobeerd aan te geven voor welke CBS-thema's de nieuwe databronnen gebruikt zouden kunnen worden.

2.1.1 Bedrijfsadministraties

Bedrijfsadministraties bevatten gegevens die voor het CBS erg interessant zijn. Het ontsluiten van de gegevens uit bedrijfsadministraties voor statistisch gebruik is een onderwerp dat veel bestudeerd wordt (Snijkers et al. 2003). De recente overheidsbrede inzet van XBRL (eXtensible Business Reporting Language) om financiële gegevens tussen overheidsinstellingen en het bedrijfsleven uit te wisselen biedt nieuwe mogelijkheden voor het CBS (Daas en Roos 2008). Het eerste resultaat van het gebruik van XBRL is de Nederlandse Taxonomie (Pasmooij en de Haas 2005). De Nederlandse 'taxonomie' bevat de gecombineerde metadata van bepaalde rapportages voor de Belastingdienst, de Kamers van Koophandel en het CBS. Hierdoor kunnen met XBRL zowel gegevens naar de Belastingdienst, de Kamers van Koophandel en het CBS worden verstuurd. Dit maakt het voor bedrijven steeds aantrekkelijker XBRL-enabled bedrijfsadministratie software te gebruiken (figuur 2). De verwachting is dan ook dat in de nabije toekomst steeds meer administraties van bedrijven gegevens in het XBRL-formaat kunnen in- en exporteren.



Figuur 1 Het gebruik van XBRL voor de uitwisseling van gegevens tussen bedrijfsadministraties

Voor het CBS wordt het dan ook interessant andere gegevens dan de gegevens die op dit moment in het CBS-deel van de Nederlandse Taxonomie zijn vastgelegd¹ m.b.v. XBRL te verzamelen. Eén van de potentieel meest interessante gegevens voor het CBS zijn die van de goederenhandel tussen de lidstaten van de Europese Unie; de zogenaamde Intrastat-gegevens. De administratieve lastendruk voor het verzamelen van de Intrastat-gegevens is namelijk erg hoog. Bijna 60% van de totale

¹ In de meest recente versie van de Nederlandse Taxonomie (versie 2.1) zijn de Productie- en de Korte termijn statistieken en de statistiek van de Investerings opgenomen.

administratieve lastendruk van het CBS op het bedrijfsleven wordt hierdoor veroorzaakt (Muller et al. 2007). De Intrastat-gegevens zijn voor een groot deel in bedrijfsadministratie(s) te vinden. Tot enige tijd geleden was het echter niet goed mogelijk de metadata van de Intrastat-uitvraag in XBRL te definiëren. Met name de uitgebreide goederenspecificatie ('de productcatalogus') was hier de oorzaak van. Daarbij komt dat voor Intrastat gegevens over individuele transacties gerapporteerd moeten worden terwijl XBRL oorspronkelijk is opgezet om over geaggregeerde financiële transacties te rapporteren (Hoffman 2006).

Recente uitbreidingen van de XBRL-specificatie maken het mogelijk beide problemen op te lossen. Het probleem van de individuele transacties kan worden opgelost door XBRL 'Global Ledger' (XBRL-GL) te gebruiken. XBRL-GL is specifiek gericht op het registreren van de afzonderlijke transacties in de boekhoudsystemen van bedrijven (XBRL 2008). Het andere probleem, de lastige te coderen goederenlijst, kan door XBRL-dimensies worden opgelost. Door deze uitbreiding is het mogelijk complexe (hiërarchische) codelijsten te definiëren (Burger, 2009). Nieuw is de combinatie van beide XBRL-uitbreidingen. Dit vormt een uitdaging die zowel voor de XBRL-gemeenschap als voor andere statistische bureaus interessant is. Door het CBS is voorgesteld dit onderzoek in het kader van Modernisation of European Enterprise and Trade Statistics (MEETS) uit te voeren. De XBRL-oplossing is overigens een alternatief voor de XML-schema aanpak die Eurostat voorstelt voor de Intrastat-opgaven (InstatXML 2003).

Dezelfde techniek kan ook gebruikt worden om een volledige uitvraag voor de Productiestatistiek te realiseren. Tot nu toe is in de Nederlandse Taxonomie alleen de generieke informatie opgenomen, maar het is met XBRL-dimensies mogelijk taxonomieën te voorzien van specificatievariabelen. Deze kunnen eventueel zelfs geïndividualiseerd worden aangeboden. Om dit laatste mogelijk te maken zou ook onderzocht moeten worden via welk kanaal deze dan moeten worden aangeboden.

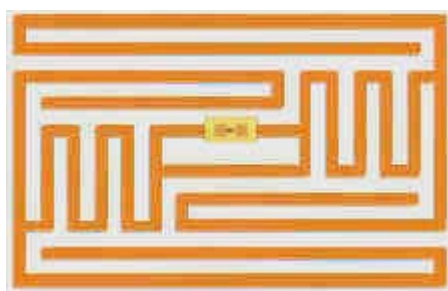
Tot nog toe werd alleen aan kleine en middelgrote bedrijven toegestaan om gegevens in XBRL-formaat aan te leveren. Het is waarschijnlijk dat ook grotere bedrijven in staat zullen zijn XBRL-gegevens aan te bieden. In samenwerking met het CBS-project voor een gecoördineerde waarneming van grote ondernemingen zal onderzocht kunnen worden hoe omgegaan wordt met deze informatie, die immers niet altijd betrekking hoeft te hebben op dezelfde eenheid als waarover informatie verlangd wordt.

2.1.2 Radio frequency identification tags

Identificatie met radiogolven (Radio frequency identification; RFID) is een technologie die gebruikt wordt om objecten van een afstand te herkennen. Dit gebeurt m.b.v. zogenaamde RFID-'tags' die op of in de te identificeren objecten zitten (figuur 3). RFID-tags kunnen 'actief' of 'passief' zijn. Actieve RFID-tags hebben een eigen energiebron. De informatie die in actieve tags is opgeslagen kan van afstand, m.b.v. een 'remote tranceiver', door radiogolven worden gelezen én geschreven. Passieve RFID-tags hebben geen eigen energiebron. Ze kunnen

uitsluitend van afstand worden gelezen. De informatie die is opgeslagen wordt verzonden met behulp van de energie van de ontvangen radiogolven. Meestal bevat het geheugen van een dergelijke RFID-tag een uniek willekeurig identificatienummer, ook wel een GUID genoemd. Daarnaast kunnen RFID-tags ook verschillen in de frequentie van de radiogolven die ze uitzenden en ontvangen. In het algemeen geldt: hoe hoger de frequentie, hoe verder het leesbereik. Nadeel van hogere frequentie RFID-tags is dat het leesbereik door metaal en vocht verminderd wordt.

De grote voordelen van het gebruik van RFID-tags zijn dat er geen fysiek contact of zichtlijn nodig is om de informatie te lezen, dat ze leesbaar zijn op relatief grote afstand en dat enkele honderden tags binnen enkele seconden gelezen kunnen worden. Ten opzichte van de streepjescode zijn dat aanzienlijke voordelen. Nadelen



Figuur 2 Een passieve RFID-tag. In het midden is de chip te zien waarin de unieke code is opgeslagen. De andere structuur is de antenne.

zijn dat de schrijfmogelijkheden van de actieve RFID-tag het mogelijk maken dat er ongemerkt fraude wordt gepleegd, er door het grote (zend/ontvang)bereik verstoring kan optreden en dat veel informatiesystemen (nog) niet ingericht zijn om de gegevens van grote hoeveelheden RFID-tags te verwerken. Voor dit laatste probleem zijn specifieke systemen (middleware en edgeware) ontwikkeld. In de praktijk worden RFID-tags momenteel voornamelijk gebruikt om de verplaatsing van producten binnen bedrijven te detecteren.

Het is voor het CBS zeer interessant de verplaatsingsgegevens van goederen die met een RFID-tag gemerkt zijn te verzamelen. RFID-tags worden veel gebruikt op (pallets van) producten die door grote supermarktketens en warenhuizen worden verkocht. Wat dat betreft zou de detailhandel een erg interessante bron van informatie kunnen zijn. Voor het gebruik van RFID-gegevens is het belangrijk dat het CBS de mogelijkheid krijgt om de verplaatsingsgegevens uit de daarvoor gebruikte systemen bij bedrijven te verzamelen. Dit biedt de mogelijkheid het transport van (een deel van de) goederen gedetailleerd in kaart te brengen zonder additionele administratieve lasten te veroorzaken. Dit onderwerp hangt nauw samen met de verplaatsingsgegevens die in de volgende paragraaf worden besproken.

De informatie die in RFID-tags is opgeslagen moet over de hele keten gestandaardiseerd zijn om door de verschillende systemen in de keten gebruikt te kunnen worden. De gestandaardiseerde gegevens (zoals productbeschrijving,

overslagpunten, etc.) zijn daardoor eveneens voor het CBS interessant, zeker wanneer die gegevens ook door andere systemen gebruikt worden. Het CBS kan signaleren aan bijvoorbeeld Eurostat of andere standaardisatie instellingen (zoals de ISO-groep) wanneer verschillende standaarden elkaar kunnen versterken, eenzelfde domein bestrijken of met elkaar in conflict zijn.

Bij de douane en Belastingdienst kan aangesloten worden op het ITAIDE project (Information Technology for Adoption and Intelligent Design). Dit project voert onder meer pilots uit met bedrijven om de import en exportstromen meer transparant te maken (ITAIDE 2008).

2.1.3 Satelliet gebaseerde navigatiesystemen

Veel auto's en vrachtauto's zijn tegenwoordig uitgerust met een satelliet gebaseerd navigatiesysteem. Een dergelijk systeem is in staat zijn positie op de aarde m.b.v. satellietgegevens bij benadering te bepalen. Een satellietnavigatiesysteem wordt daarom ook vaak aangeduid als een 'Global Positioning System' (GPS). In de volksmond wordt een dergelijk systeem ook wel een 'Tomtom' genoemd.

De GPS-gegevens van vrachtauto's zijn voor het CBS een zeer interessante bron van informatie. Ze bevatten namelijk informatie die relevant is voor de vervoersstatistiek. Momenteel worden voor de vervoersstatistiek gegevens van enquêtes en (zeer beperkt) XML-gegevens uit de administratie van beroepsvervoerders gebruikt. De vervoerders die in de steekproef vallen moeten op een gedetailleerd niveau gegevens leveren over de verplaatsing en de lading van hun vrachtwagens. Het deel van de gegevens dat betrekking heeft op de verplaatsing van de vrachtwagens is vaak al elektronisch beschikbaar. Veel van de vervoerders bezitten namelijk vrachtwagens die zijn uitgerust met telemetriesystemen. Deze systemen slaan, onder andere met behulp van GPS-signalen, gegevens op over de positie, verplaatsing en snelheid van de vrachtwagen (zie figuur 3). Op gezette tijden ontvangen de vervoerders deze gegevens met behulp van een digitale verbinding. Het CBS kan deze gegevens gebruiken om de verplaatsing te meten van de vrachtwagens die in het wagenpark van het geënquêteerde bedrijf vallen. Het bedrijf zou dan alleen nog de gegevens over de lading moeten verstrekken. Te verwachten valt dat het gebruik van GPS-gegevens tot een aanzienlijke kwaliteitsverbetering zal leiden.



Figuur 3 Illustratie van het gebruik van GPS-gegevens voor het bepalen van de afgelegde route (getoond op een Google-maps afbeelding).

Het CBS ontvangt dan immers van die voertuigen in de steekproef de volledige set van verplaatsingsgegevens; dat wil zeggen het aantal en de route van alle gereden ritten. Voor het mogelijke gebruik van GPS-gegevens door het CBS is nog aanvullend onderzoek noodzakelijk. Dit betreft de volgende onderwerpen:

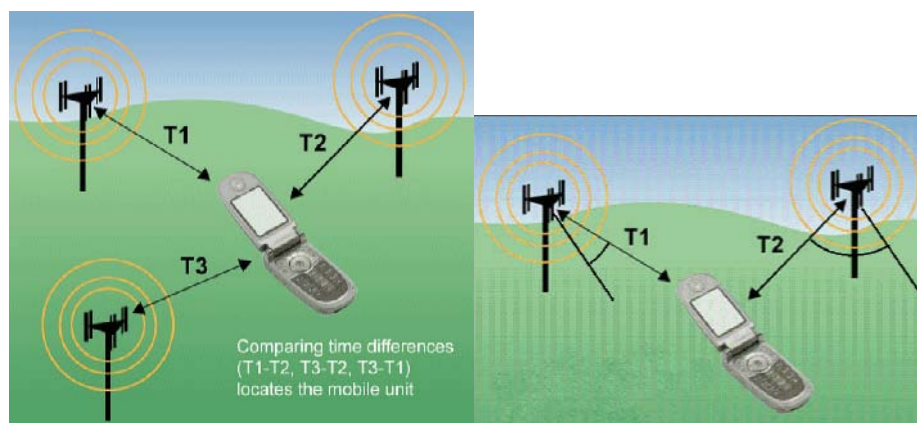
- a) Onderzoek naar de standaarden die gebruikt worden voor het opslaan en doorsturen van GPS-gegevens door de systemen die vervoersbedrijven gebruiken. Een eerste verkenning laat zien dat er waarschijnlijk niet één standaard is. Bedrijven bieden verschillende ‘vehicle tracking’ systemen aan die ieder een ander rapportformaat lijken te genereren. Nader onderzoek dient uit te wijzen of het aantal systemen zodanig beperkt is dat het voor het CBS rendabel is om de gegevens als zodanig af te nemen. Als het merendeel van de bedrijven slechts een of twee verschillende type systemen gebruikt dan kan het rendabel zijn koppelingen te realiseren. Als het om een veelvoud van dergelijke systemen gaat, wordt het opzetten en onderhouden van dergelijke koppelingen kostbaar en daardoor minder interessant. Mogelijk is het GPX-formaat, een voorstel voor een standaard uitwisselingsformaat van GPS-gegevens, hiervoor een oplossing (GPX 2008).
- b) Onderzoeken van de standaarden op structuur en bruikbaarheid van de gegevens. De gegevens zullen waarschijnlijk niet direct gebruikt kunnen worden. Gewenste gegevens over het verreden aantal kilometers, grensovergangen en laad- en lospunten moeten afgeleid worden uit een set van gegevens van tijdseries van gekoppelde GPS-coördinaten.

Voor het daadwerkelijk kunnen beoordelen van de bruikbaarheid van de gegevens is de medewerking van een aantal transportbedrijven noodzakelijk. Van één of meer bedrijven zullen de opgeslagen GPS-gegevens beschikbaar moeten worden gesteld. Deze gegevens zullen vergeleken moeten worden met de verwachte set van gegevens (enquête of XML). Indien de GPS-gegevens, na een eventuele bewerking, overeenkomen met de verwachting zal in een grotere praktijktest de bruikbaarheid van de GPS-gegevens bepaald moeten worden.

Een andere mogelijke bron van data over verkeersstromen is het rekeningrijdensysteem. Deze data zal waarschijnlijk gekoppeld zijn aan kentekens en zal dus als een zeer bruikbare bron kunnen dienen om verkeersstromen in kaart te brengen.

2.1.4 Mobiele telefoons

Veel Nederlanders bezitten een mobiele telefoon (GSM). In 2007 was dit bij ruim 90% van de huishoudens in Nederland het geval (CBS 2008). Naast het belgedrag levert een mobiele telefoon -mits die aan staat- nog andere interessante informatie op, namelijk locatie- en verplaatsingsgegevens. Mobiele telefoonaanbieders zijn in staat de locatie van de telefoon uit de contactgegevens van de telefoon met verschillende telefoonmasten (GSM-masten) te destilleren (Tweede Kamer 2007). Er zijn verschillende technieken om dit te doen. De locatie kan vrij ongedetailleerd worden achterhaald door simpelweg te kijken bij welke GSM-mast een mobiele telefoon is aangemeld. Ook bestaan er meer geavanceerde technieken om ervoor te zorgen dat de locatie van een mobiele telefoon veel nauwkeuriger kan worden achterhaald. Een belangrijke techniek is de 'Time Advance' techniek, waarbij de tijd wordt gemeten die een signaal nodig heeft om van de mobiele telefoon naar de GSM-mast te komen. Voor een optimale lokalisering zijn de gegevens van drie verschillende GSM-masten noodzakelijk, maar met minder kan het ook (figuur 4). Een voorbeeld van het gebruik van deze techniek is de lokalisatiemogelijkheid die enkele bedrijven in Europa zeer recent zijn gaan aanbieden. Deze bedrijven geven abonneementhouders tegen betaling inzicht in de locatie van bepaalde mobiele telefoons. Deze service wordt o.a. aangeboden aan ouders die zo, op elk gewenst moment tot maximaal een aantal keren per dag, kunnen zien waar de mobiele telefoon van hun kind zich bevindt (Telecomwereld 2008, Koning 2008).



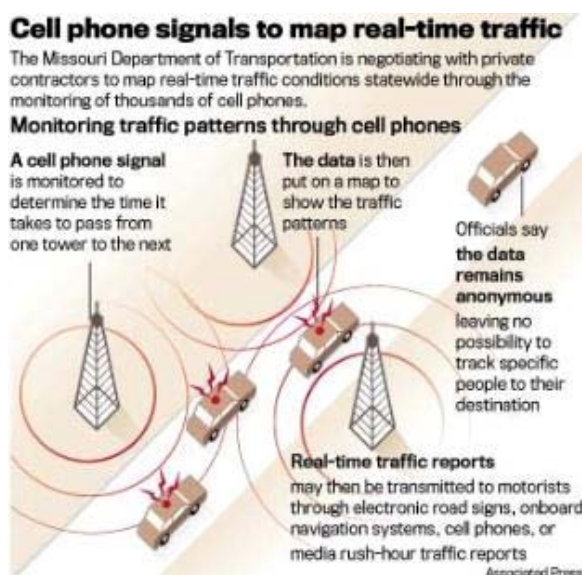
Figuur 4 Locatiebepaling van een mobiele telefoon met de connectiegegevens van drie en twee GSM-masten. Met drie masten zijn de tijd tussen het verzenden en ontvangen van de signalen voldoende. Bij twee masten moet tevens de richting van verbinding mee worden genomen.

Het gebruik van de locatie-informatie van een mobiele telefoon voor andere activiteiten wordt 'location based services' genoemd. Deze services zijn erop gericht

om, op basis van de positie van de mobiele telefoon van een klant, bepaalde specifieke acties te ondernemen. Enkele voorbeelden illustreren dit gebruik. Wanneer iemand door een luchthaven loopt om voor zijn/haar vlucht te gaan inchecken, kan de aanwezigheid van de mobiele telefoon van die persoon op de luchthaven al voldoende zijn om deze handeling te verrichten. Technisch gezien is dit haalbaar, of iemand op een dergelijke dienst zit te wachten is een andere vraag. Een ander voorbeeld is de fileservice die Vodafone en Tomtom sinds enige tijd samen aanbieden. Vodafone levert hierbij file-informatie aan gebruikers van een bepaald model routeplanner van Tomtom (Webwereld 2007). De gegevens zijn gebaseerd op de locatie van de Tomtom en o.a. de locatie-informatie van andere mobiele telefoons met een Vodafone abonnement in Nederland (figuur 5). Op basis van deze zeer actuele informatie kan de Tomtom vervolgens aangeven of er files op de geplande route voorkomen en indien nodig direct een alternatieve route voorstellen.

Omdat veel Nederlanders een mobiele telefoon bezitten en deze vaak continu aan hebben staan kunnen de locatiegegevens een zeer interessante bron van informatie voor het CBS zijn. Op deze manier kan namelijk informatie verkregen worden over het verplaatsingsgedrag van de Nederlanders met een mobiele telefoon. Bij een dergelijk onderzoek zal zorgvuldigheid voorop moeten staan. Belangrijk hierbij is de afweging hoe de gegevens afkomstig van mobiele telefoons en mobiele telefonie zendmasten zijn te gebruiken zonder dat ze te herleiden zijn naar individuele personen.

De vraag hoe we ons verplaatsen kan voor de meest uiteenlopende vragen van belang zijn. Dit kan interessant zijn voor het bepalen van woon-/werk verkeer; hoe is het verloop van het woon-/werk verkeer gedurende de dag? Ook zou het CBS op basis van deze gegevens nieuwe, meer populaire, statistieken kunnen uitbrengen.



Figuur 5 Gebruik van locatiegegevens van mobiele telefoons voor file-informatie.

Denk aan bezoekersaantallen van winkelcentra en dergelijke. In relatie tot het toenemende milieubewustzijn zijn de verplaatsings- en locatiegegevens eveneens interessant. Zo zouden met behulp van de GSM-gegevens verkeersintensiteiten en gereden autokilometers naar wegtype bepaald kunnen worden. Dit is o.a. interessant voor verkeersveiligheidsonderzoek en de bepaling van de uitstoot van CO₂ en fijnstof binnen de bebouwde kom.

Naast het feit dat mobiele telefoonaanbieders in staat zijn om op een accurate manier de positie van hun klanten te registreren (zie de eerdere voorbeelden) worden deze gegevens ook voor nationale veiligheidsredenen bewaard. Hierover is in januari 2008 een beslissing genomen door de Tweede kamer (Tweede Kamer 2008). Er is een minimale periode van 18 maanden vastgesteld voor het bewaren van de locatie- en telefoonverkeergegevens. Het resultaat van de samenwerking tussen Tomtom en Vodafone, laat zien dat het op zich ook geen probleem is om 'realtime' gebruik te maken van de locatiegegevens van veel gebruikers.

Een belangrijk uitgangspunt voor het gebruik van mobiele telefoongegevens door het CBS is het waarborgen van de privacy van individuele personen. In principe zijn de locatiegegevens van mobiele telefoons op vier verschillende detailniveaus door het CBS te gebruiken. Deze zijn in tabel 1 als alternatieve A t/m D opgenomen. De vier voorgestelde alternatieven zijn de vier meest logische keuzen die gemaakt kunnen worden wanneer naar de privacygevoeligheid en het detailniveau van de informatie gekeken wordt. De alternatieven zijn in tabel 1 gescoord op vier voor het CBS zeer relevante kwaliteitsaspecten van de eindresultaten. Om schijnnaauwkeurigheden te voorkomen worden in de tabel slechts een '+' en een '-' gebruikt om een voor- dan wel nadeel aan te duiden. Daarnaast wordt er in de tabel nog een ander symbool gebruikt, namelijk '/'. Dit symbool geeft aan dat voor een bepaald aspect niet eenduidig te bepalen valt of dit een voor- of nadeel is. Zo wordt het '/' symbool bij alternatief B gebruikt om aan te geven dat de voor- en nadelen van privacy en stratificeerbaarheid erg variabel zijn. In principe is het namelijk mogelijk om van deelpopulaties meer (detail)informatie te vragen, wat dan alsnog

Tabel 1. Overzicht van de voor- en nadelen van vier verschillende detailniveaus bij het gebruik van locatiegegevens van mobiele telefoons.

		privacy	detailniveau	stratificeerbaarheid	hoeveelheid
A	volledige aanlevering van telefoonnummers met posities	-	+	+	-
B	aanlevering van SIM-nummers met posities, of aanlevering van een ongerelateerd nummer met posities	/	+	/	-
C	aanlevering van posities	+	+	-	-
D	aanlevering van aantal toestellen binnen een bepaald gebied (bijvoorbeeld postcode-gebied)	+	-	-	+

een voordeel oplevert. Daarnaast is er een verschil tussen de data-analyse eigenschappen van de alternatieven A, B en C en het alternatief D. Bij alternatief D worden uitsluitend *netto verplaatsingen* bepaald. Dus: stel er gaan meer mensen van

Heerlen naar Maastricht tijdens de ochtendspits dan van Maastricht naar Heerlen, dan zal er bij alternatief D slechts een trage stroom van Heerlen naar Maastricht te zien zijn. Bij alternatieven A, B en C is het wel mogelijk de afzonderlijke mobiele telefoons (als 'deeltjes') waar te nemen. Daarbij is dus te zien hoe de individuele mobieltjes zich door Nederland verplaatsen. Bij alternatief C is er sprake van een modelaanne. Het is hierbij namelijk niet mogelijk een mobiele telefoon uniek te identificeren. Alternatief C beschrijft dus uitsluitend de manier waarop telefoons zich verplaatsen. Bij alternatief C is waarschijnlijk een hogere sample-rate noodzakelijk om de samenhang tussen deeltjes nauwkeurig genoeg te kunnen berekenen. In het geval van integrale waarneming vervallen echter veel van de eerder genoemde problemen. Bij alternatieven A en B is het wel mogelijk de telefoons direct te volgen. Bij model B is het echter niet mogelijk de gebruiker van de mobiele telefoon te identificeren, bij model A is dat wel het geval. Het zal duidelijk zijn dat bij onderzoek naar alternatief A de privacygevoeligheid van de gegevens zal leiden tot vraagtekens over het waarborgen van de privacy. Maatregelen om dit te waarborgen zijn bij dit onderzoek noodzakelijk.

Als alternatief zou bij model B i.p.v. het SIM-nummer, het IMEI-nummer van het telefoontoestel kunnen worden aangeleverd. Het IMEI-nummer is een 14-cijferig nummer waarmee een GSM-toestel uniek wordt geïdentificeerd. Hierdoor worden de gegevens gekoppeld aan een toestel i.p.v. aan een SIM-kaart. Het nummer van een SIM-kaart is namelijk uniek voor de *kaart* die in een GSM moet worden gestoken om van een bepaald telefoonnetwerk gebruik te kunnen maken. Een SIM-kaart kan dus in een ander toestel worden geplaatst. Voordeel van het SIM-nummer is wel dat deze (vrijwel altijd) persoonsgebonden is omdat dit nummer met een abonnement of een prepaid aansluiting samenhangt. Ook hier zijn zorgen over de privacy op hun plek, omdat op basis van de positie in sommige gevallen ook de identiteit van een eigenaar van de mobiele telefoon herleidbaar kan zijn. Ook hier is zorgvuldigheid een noodzakelijke vereiste.

De onvolledige dekking van de Nederlandse bevolking door de populatie die een mobiele telefoon bezit, het feit dat sommige personen meer dan één toestel bezitten (b.v. werk en privé) en een mogelijk verschil tussen de eigenschappen van personen met en zonder een mobiele telefoon zullen de verkregen resultaten beïnvloeden. Hieronder worden enkele gevolgen voor de ophoging van gemeten resultaten besproken. De vraag is of het mogelijk is op een eenvoudige manier (zonder te veel kosten) de eigenschappen te kunnen achterhalen van mensen met en zonder een mobiele telefoon. Zonder additionele informatie is dit alleen bij alternatief A en misschien ook bij alternatief B mogelijk voor het CBS. Indien voor alternatief C of D wordt gekozen zullen de providers zeer waarschijnlijk subpopulaties moeten aanleveren. Daarnaast dient ook met de mogelijke gevolgen van personen met meerdere mobiele telefoons rekening te worden gehouden. Wat dat betreft biedt alternatief A de meeste mogelijkheden om dekkingsproblemen op te lossen. Een andere nuttige aanvulling is om bij één of meerdere grote CBS-enquêtes het mobiele telefoonnummer(s) van respondenten te vragen. Hiermee zijn dan voor een deel van de populatie belangrijke achtergrondkenmerken beschikbaar. Daarna is het

noodzakelijk de nummers in de grote hoeveelheid GSM-data te identificeren. Uiteraard zal met deze gegevens zorgvuldig om moeten worden gegaan.

Ook de hoeveelheid data speelt een rol in het beoordelen van de criteria. Hoe hoger het detailniveau, hoe meer data verwerkt en geanalyseerd moeten worden. De hoeveelheid data kan bij alternatief A en B in zowel het aantal records als de fysieke ruimte op schijf enkele ordes groter uitvallen dan waar het CBS ervaring mee heeft.

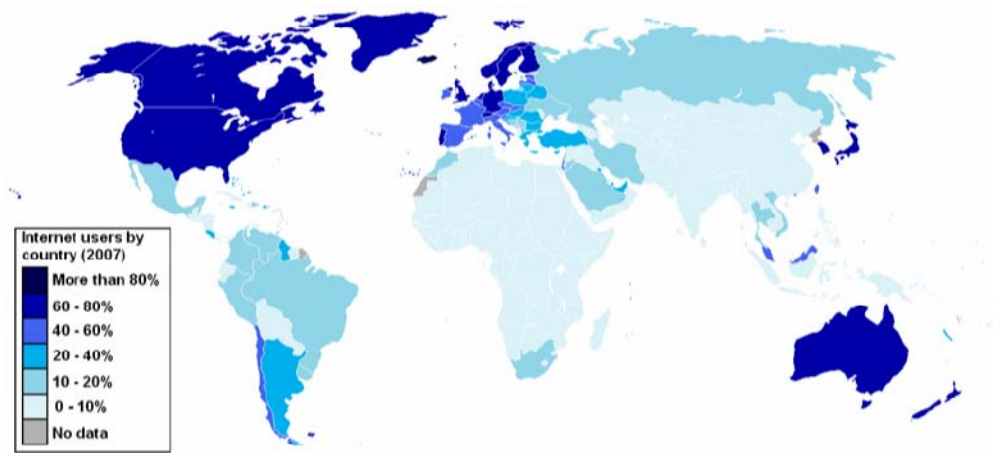
Voor het mogelijke gebruik van de locatiegegevens van mobiele telefoons door het CBS is nog aanvullend onderzoek noodzakelijk. Dit betreft de volgende werkzaamheden:

- a) Contacten leggen met mobiele telefoonaanbieders om data te verkrijgen.
- b) Optimaliseren van de keuze tussen het gewenste detailniveau en de privacyaspecten van de gebruikers
- c) Nadere studie van de technische aspecten. Deze hebben o.a. betrekking op het dataformaat, eventueel noodzakelijke conversies, beveiligingsaspecten en hoe met de mogelijk grote hoeveelheden gegevens moet worden omgegaan
- d) Bruikbaarheid van de resultaten en het mogelijk gebruik in statistische publicaties.

Het zou interessant zijn een dergelijk onderzoek in Europees verband uit te voeren. Gezien de pioniersrol van Finland op het gebied van mobiele telefonie zou bij voorkeur met het statistische bureau van dat land samengewerkt moeten worden. Voordeel van de onderzoekers van het Finse Statistische Bureau is dat deze reeds veel kennis in huis hebben en dat ze beschikken over een database met daarin alle Finse mobiele telefoonnummers.

2.1.5 Internet

Het internet is een wereldwijd netwerk van onderling verbonden computernetwerken. Het vindt zijn oorsprong in 1969 toen steeds meer computers onderling werden verbonden. In die tijd ontstonden de eerste (lokale) computernetwerken. Deze netwerken werden min of meer tegelijkertijd door zowel het Amerikaans leger als door een aantal Amerikaanse universiteiten opgericht. In de loop van de tijd werden steeds meer computers en computernetwerken op andere netwerken aangesloten. Aldus ontstond het internet; een afkorting van Interconnected Networks. Inmiddels is het internet een wereldomvattend fenomeen dat het karakter van een massamedium heeft gekregen. Het is het meest gebruikte communicatiemiddel ooit. Figuur 6 geeft een overzicht van het percentage internetgebruikers in verschillende landen in de wereld in het jaar 2007. In Nederland had in dat jaar 83% van de huishoudens toegang tot internet. Bijna driekwart van de huishoudens beschikte zelfs over een snelle (breedband) verbinding (CBS 2008).



Figuur 6 Percentages internetgebruikers per land in 2007.

In het dagelijkse spraakgebruik is het internet vaak een synoniem voor het World Wide Web. Dit is slechts één van de vele diensten die via het internet kunnen worden gebruikt. Andere bekende diensten zijn e-mail, FTP en usenet (nieuws groepen). De laatste jaren wordt het internet ook steeds meer gebruikt voor het ‘streamen’ van audio en video, het uitwisselen van bestanden via peer-to-peer netwerken (zoals BitTorrent, eDonkey en Gnutella), telefonie (Voice Over IP) en tekstgebaseerde communicatie (chatten; o.a. MSN of ICQ).

Op het internet is niet alleen zeer veel informatie te vinden, het is zelf ook een interessant maatschappelijk en economisch fenomeen (Jones 1998). Voor het CBS kan het internet een zeer belangrijke (aanvullende) bron van informatie zijn (Dialogic et al. 2008, Bethlehem 2008). In de jaarlijkse publicatie “De digitale economie” geeft het CBS reeds een overzicht van de digitalisering van de Nederlandse samenleving (CBS 2008). Naast de onderwerpen die in deze publicatie besproken worden en het afnemen van enquêtes via internet, zijn er een aantal onderwerpen te bedenken waarover het CBS -m.b.v. het internet- gegevens zou kunnen verzamelen. Dit zijn:

- Internetdata in Nederland
 - o De hoeveelheid
 - o De gevolgde route (de ‘flow’)
 - o De inhoud van de data
- Sociale en economische activiteiten
 - o Persoonsgegevens
 - o Opinies in weblogs
 - o Sociale netwerken
 - o Bedrijfsinformatie
 - o Digitale economie

In de volgende paragrafen worden deze onderwerpen en het mogelijk gebruik nader toegelicht. Veel van de gegevens op het internet, zeker die op Word Wide Web-pagina’s zijn publiekelijk beschikbaar. Vanuit dat standpunt bekeken zullen hierbij minder privacy-gevoelige problemen spelen. Dit doet echter niets af aan het feit dat het CBS ook met deze gegevens zorgvuldig om zal moeten gaan. Omdat het internet

steeds in ontwikkeling is, valt te verwachten dat er in de loop van de tijd steeds weer nieuwe informatie over voor het CBS -mogelijk- interessante onderwerpen te vinden is. Het CBS zal het internet dus steeds goed in de gaten moeten houden.

2.1.5.1 Internetdata

Veel internetproviders houden het 'dataverbruik' van hun abonneenthouders in de gaten. Onder dataverbruik wordt de combinatie van het down- en uploadverkeer verstaan. Dit wordt gedaan omdat er, vooral bij goedkope abonnementen, vaak sprake is van een datalimiet. Het is dan niet toegestaan meer dan een bepaald aantal gigabyte per maand aan data te down- en uploaden. Met name kabelmaatschappijen (bijvoorbeeld UPC en @Home) schijnen dit goed in de gaten te houden omdat de kabel een gedeeld medium is. Meerdere gebruikers delen hierbij dezelfde capaciteit. Bij andere type internetverbindingen (zoals ADSL) is er van delen veel minder sprake. Bij het meten van het dataverbruik wordt voor elke abonnee (feitelijk: elk IP-adres²) bijgehouden hoeveel gigabyte er gedown- of geupload wordt. Het is niet bekend of er ook in de gaten wordt gehouden naar welk IP-adres de data verstuurd wordt dan wel van welk adres deze afkomstig is.

Bij het monitoren van het dataverkeer kan in eerste instantie gedacht worden aan het monitoren van de drukte op internet; analoog aan het bijhouden van verkeersdrukke op het wegennet in Nederland. Hierbij kan b.v. het internetverkeer op een belangrijk knooppunt in Nederland in de gaten worden gehouden. Het belangrijkste knooppunt van Nederland bevindt zich in Amsterdam; de Amsterdam Internet Exchange. Het is echter belangrijk te beseffen dat op het internet:

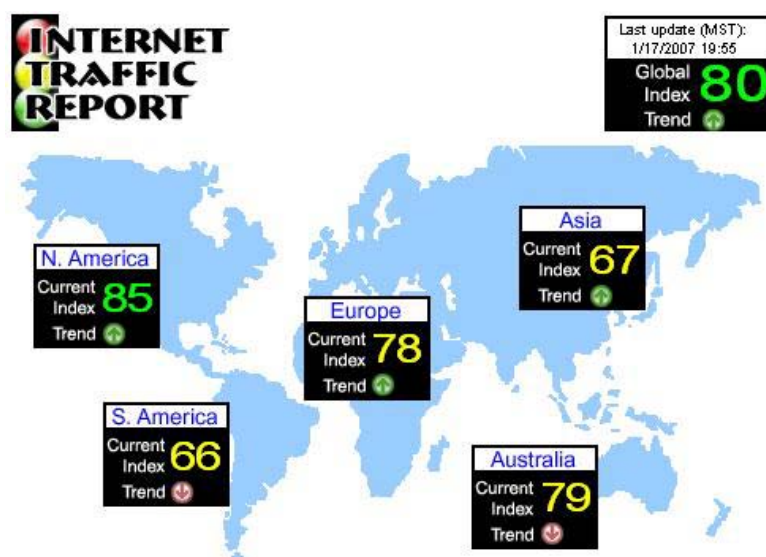
- Alle data bij het versturen in kleine pakketten wordt opgesplitst
- De pakketjes niet allemaal dezelfde route tussen afzender en ontvanger af hoeven te leggen
- De gebruikte applicatie bepaalt of het afleveren (de aankomst) van de pakketjes gegarandeerd wordt
- De snelheid waarmee de pakketjes zich verplaatsen sterk door het andere internetverkeer wordt beïnvloed.

Er is een website die een actueel overzicht geeft van de huidige drukte op het internet. Deze is te vinden op <http://www.internettrafficreport.com/>. Op deze website wordt de drukte van de belangrijkste internetknooppunten op de verschillende werelddelen (m.u.v. Afrika en Antarctica) in een indexscore (de z.g. 'traffic' index³)

² Op het internet wordt voor de onderlinge communicatie tussen computers, in het algemeen, het Internetprotocol (IP) gebruikt. Een IP-nummer identificeert een internetabonnee. Omdat sommige providers hun gebruikers nogal eens van IP-adres laten wisselen (z.g. dynamische adressen) maakt een abonneethouder niet altijd van hetzelfde IP-adres gebruik. In bijlage 3 wordt het IP in meer detail besproken

³ De "traffic index" is een getal tussen de 0 en de 100 waarbij 0 langzaam en 100 snel is. De score wordt bepaald door de huidige snelheid waarmee een knooppunt reageert te vergelijken met de snelheden van reageren in de afgelopen week. De snelheid wordt bepaald door de tijd

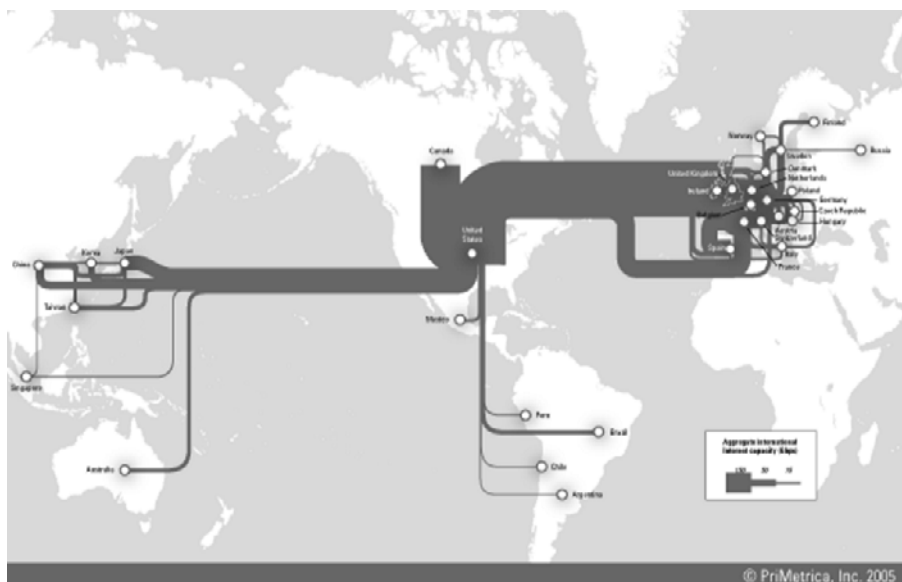
en een trend (toename, afname, geen verandering) weergegeven (figuur 7). De drukte kan tot op landenknooppunt gedetailleerd worden bekeken. Voor elk knooppunt is, voor een periode van 24 uur, de traffic index, de responsetijd en eventuele pakketverlies in een grafiek te zien. De gegevens worden om de 5 minuten ververs. Voor Nederland is het meest gedetailleerde niveau de activiteit van het knooppunt in Amsterdam.



Figuur 7 De drukte op internet zoals bepaald door de traffic index en trend.

Daarnaast is het ook mogelijk de route die de datapakketten afleggen (de 'flow') over het internet te bestuderen. Het datatransport over het internet in 2005 is in figuur 8 weergegeven. Doordat de pakketten, tussen de afzender en ontvanger, verschillende routes over het internet kunnen afleggen, is het lastig om de 'flow' zeer gedetailleerd te bepalen. Hierover zijn een aantal wetenschappelijke publicatie verschenen (Cleveland en Sun 2000, Di Cairano-Gilfedder en Clegg 2005). De complexiteit (en heterogeniteit) van de structuur van het internet, de informatie in de datapakketten en de verdeling in de aankomst van de verstuurd pakketten maken het lastig het transport nauwkeurig te bepalen. Dit onderwerp heeft een interessante analogie met het volgen van de verspreiding van virussen over het internet (Moreno et al. 2002).

te meten tussen het verzenden en ontvangen van een 'ping'-signaal naar het betreffende knooppunt.



Figuur 8 Weergave van de ‘flow’ van het internetverkeer over de wereld in 2005

Het is ook mogelijk naar de inhoud van het dataverkeer op het internet te kijken. Vaak richt het bestaande onderzoek zich op het bepalen van het aandeel e-mail, videogerelateerde gebruik en/of het illegaal uitwisselen van bestanden (bijvoorbeeld via P2P) in het totale dataverkeer (Adams 2005, Ipoque 2006, 2007). Het nauwkeurig bepalen van het aandeel van de verschillende soorten internetverkeer is niet altijd even eenvoudig (Ipoque 2006, Dialogic et al. 2008).

Naast de drukte, het transport en de inhoud van de data over het internet zijn er nog veel meer karakteristieken te bepalen, voorbeelden hiervan zijn capaciteiten, routeringen en pakketverlies. Ook de aantallen bezoekers van bepaalde (categorieën van) websites en het aandeel van de verschillende ‘browsers’ die voor websurfen gebruikt worden behoren tot de groep van internetdata gerelateerde statistieken. Uitgebreide overzichten van de vele mogelijke internetkarakteristieken zijn te vinden in Monk en Claffy (1996), Internet statistics guide (2002) en CAIDA (2008).

Hoe groter de invloed van het internet op het sociaal en economisch leven in Nederland hoe groter de behoefte zal zijn hier, bij voorkeur door een onafhankelijk en betrouwbaar instituut (zoals het CBS), cijfers over te gaan publiceren. Methoden om aan deze gegevens te komen zijn het monitoren van de verschillende karakteristieken op een belangrijk internetknooppunt, bij een aantal grote internetproviders en/of bij een steekproef van gebruikers. In de eerste twee gevallen zal het CBS moeten onderhandelen om van deze gegevens gebruik te kunnen maken. Voor het verkrijgen van gegevens van gebruikers zou het installeren van een programma dat het computer- en internetgebruik continue in de gaten houdt noodzakelijk zijn. Dit zou uiteraard alleen met toestemming van de gebruiker kunnen geschieden (Dialogic et al. 2008).

2.1.5.2 Sociale en economische activiteiten

Het internet bevat zeer veel webpagina's. Deze pagina's bevatten o.a. informatie over bedrijven en personen. Het deel van de webpagina's dat informatie bevat over personen kán voor het CBS een interessante aanvullende bron van informatie zijn. Het internet wordt door veel personen voor allerlei sociale activiteiten gebruikt. Voorbeelden van dergelijk gebruik zijn het schrijven en becommentariëren van een online dagboek, het gezamenlijk spelen van spellen en het 'interacteren' met anderen in een virtuele wereld. Ook het bijhouden door personen van pagina's met persoonlijke informatie (Hyves, LinkedIn) gebeurt op een steeds grotere schaal. Voor het CBS zijn dit interessante fenomenen, maar met flinke haken en ogen als het om de privacy van personen en huishoudens gaat. Kunnen en willen we als CBS op systematische wijze persoonlijke informatie van websites halen voor onderzoeks- of waarneemdoeleinden? Als dit al gewenst is, zal het waarborgen van de privacy van personen en bedrijven hoog op de agenda moeten staan bij onderzoek naar de bruikbaarheid van deze fenomenen.

Een voorbeeld van dit gebruik: heel wat personen houden een 'online dagboek' bij. Een dergelijk dagboek wordt een weblog of blog genoemd en is meestal door iedereen te lezen. Geschat wordt dat er in Nederland een klein miljoen weblogs zijn (Blogresearch 2008). Wereldwijd zijn dit er bijna 113 miljoen (Technorati 2008). De informatie in weblogs (ook wel de 'blogosphere' genoemd) kan voor het CBS interessant zijn. Via weblogs kan het CBS in de gaten houden over welke onderwerpen er veel wordt geschreven. Dit kan voor het CBS aanleiding zijn om een persbericht uit te brengen over dat onderwerp, met betrouwbare CBS-cijfers, ter ondersteuning van het debat. Daarnaast zou ook onderzocht kunnen worden of weblogs -in de toekomst- als bron of aanvullende bron van informatie over personen gebruikt zou kunnen worden. Dit dient uiteraard met de nodige voorzichtigheid aangepakt te worden. Naast de privacy gevoeligheid van gegevens is hier immers sprake van een zich zelfselecterende groep. Voorbeelden van de studie van weblogs is te vinden op de websites van de 'Annual Workshop on Weblogging' (WWE 2006).

De laatste paar jaar is er op het internet sprake van een opvallende groei van z.g. 'sociale netwerk' sites. Dit zijn websites waar gebruikers, na aanmelding, een profiel opmaken -met persoonlijke gegevens- waarbij ze tevens aan kunnen geven welke van de andere gebruikers hun vrienden zijn. Hierdoor zijn van veel personen persoonlijke gegevens online beschikbaar inclusief de relatie met anderen. Voorbeelden van 'sociale netwerk' sites zijn Facebook, LinkedIn en Hyves. Ook websites met schoolgegevens behoren tot deze groep. Een voorbeeld van een dergelijke site is Schoolbank.nl. Hierbij geeft iemand aan in welke jaren hij/zij op welke scholen heeft gezeten. Hierdoor zijn dus voor een bepaalde groep personen opleidingsgegevens online beschikbaar. Hiervan is overigens vaak slechts een beperkt deel publiekelijk beschikbaar. Meer voorbeelden van internetgerelateerde sociale studies zijn te vinden op de website van de 'Association of Internet researchers' (Internetresearchers 2008) en die van de 'Deutschen Gesellschaft für Online-Forschung' (DGOF 2008).

Het internet bevat echter ook veel bedrijfsgerelateerde gegevens. Daarmee zijn uiteraard de webpagina's van de bedrijven een belangrijke bron van informatie. Hiermee zouden bedrijfsgegevens kunnen worden gecontroleerd maar mogelijk zelfs ook verzameld. Tevens kan de informatie worden gebruikt om bedrijven beter te typeren. Indien de gewenste informatie op het internet beschikbaar is hoeft het bedrijf niet meer met een verzoek om informatie te worden lastig gevallen. Een bedrijfswebsite bevat daarnaast nog veel meer interessante informatie. Zo bevatten bedrijfswebsites vaak actuele informatie over de prijzen van de producten die ze verkopen. Ook deze informatie zou door het CBS gebruikt kunnen worden. Een voorbeeld is het verzamelen van (actuele) prijzen van producten die door supermarkten verkocht worden. Voor het verzamelen en vergelijken van de (beoogde verkoop)prijzen van huizen wordt een dergelijke aanpak sinds een aantal jaar door onderzoekers van de Vrije Universiteit van Amsterdam gebruikt (Gautier et al. 2006). Door het CBS is, in het kader van de Consumenten Prijs Index, in 2008 onderzoek uitgevoerd naar het structurele gebruik van gegevens van de website www.funda.nl.

Met behulp van het internet worden ook producten gekocht en verkocht. Over de economische activiteiten die met het internet te maken hebben publiceert het CBS reeds het een en ander, dit is beschreven in de jaarlijkse publicatie "De digitale economie" (CBS 2008). Met name de online verkregen omzet uit elektronische transacties (e-commerce) geeft een redelijke beeld van de handel en Wandel op het internet door bedrijven. Daarnaast is er op het internet echter ook sprake van economische activiteiten die niet (direct) door het CBS worden bijgehouden, denk aan het kopen en verkopen van producten op sites zoals Marktplaats en eBay e.d. Dit kan over aanzienlijk bedragen gaan, in een rapport van Dialogic et al. (2008) wordt het totale transactiebedrag van Marktplaats.nl in 2006 geschat op € 4,7 miljard. Los van de betrouwbaarheid van dit cijfer, is duidelijk dat we hier met aanzienlijke bedragen te maken hebben. Voor het CBS kan het interessant zijn over deze economische activiteiten gegevens te verzamelen en te publiceren.

Voor het verzamelen van gegevens op het internet kunnen verschillende aanpakken worden gevolgd. De meest directe is het (automatisch) verzamelen van de benodigde gegevens via de provider van de website. Dit is bijvoorbeeld toe te passen bij 'sociale netwerk' sites en op openbare koop- en verkoopsites. Een voorbeeld van een dergelijke aanpak is voor Marktplaats.nl beschreven in het rapport van Dialogic et al. (2008). Een andere aanpak, die met name interessant is voor het verzamelen van gegevens op webpagina's en in weblogs, is met behulp van z.g. 'webspiders'; ook wel 'webcrawlers' genoemd. Dit zijn computerprogramma's die het internet op een methodische en geautomatiseerde manier doorzoeken. Dergelijke programma's maken of een lokale kopie van de gevonden pagina's of slaan de tekstuele inhoud van de pagina's op zodat ze later onderzocht kunnen worden. Voor het verzamelen van prijzen is een dergelijk aanpak ook geschikt mits de gegevens als tekst

beschikbaar zijn⁴. Meer informatie over het geautomatiseerd verzamelen van internetgegevens is te vinden op <http://www.robotstxt.org/> en http://en.wikipedia.org/wiki/Internet_bot.

Bij de analyse van de grote hoeveelheden gegevens zullen vaak ‘textmining’-achtige technieken worden gebruikt. Hierbij wordt geprobeerd op een slimme manier zoveel mogelijk kennis uit vrije tekst te halen. Nadeel van dit interdisciplinaire vakgebied is dat het een relatief nieuw terrein van onderzoek is waardoor het nog volop in ontwikkeling is (Tan 1999). Ook het gebruik van de gegevens op webpagina’s heeft zo zijn moeilijkheden. De belangrijkste nadelen zijn dat:

- de betrouwbaarheid van de informatie sterk varieert per aanbieder
- de informatie onvolledig, verouderd of ongedateerd kan zijn
- de structuur van de informatie elk moment kan veranderen.

2.1.5.3 Andere internet gerelateerde onderwerpen

Tijdens het vooronderzoek dat voor dit rapport verricht is zijn, voor wat betreft de studie van het internet en internetgegevens, vele verschillende onderwerpen en toepassingen gevonden. We willen dit de lezer niet onthouden, omdat het interessante beginpunten kunnen zijn voor nader onderzoek en/of het beantwoorden van een aantal vragen. Hieronder volgt een lijstje van de interessantste, niet eerder genoemde onderwerpen. Aan elk van de onderwerpen is een internetverwijzing toegevoegd:

- De groei van het internet (http://www.firstmonday.org/issues/issue3_10/coffman/)
- Gegevens over computergebruik (<http://webstats.motigo.com/s?tab=1&link=5&id=710309>)
- Internetstatistieken over Nederland (<http://www.internetworldstats.com/eu/nl.htm>)

De volgende website geven interessante achtergrondinformatie:

- Methodologie van Web onderzoeken (<http://www.websm.org/>)
- Archivering op internet (<http://www2.internl.net/nieuws/nieuwsbrief/archief/2001/11/3.it>)

2.2 Toekomstige nieuwe databronnen

Naast de hierboven genoemde databronnen zijn er ontwikkelingen gaande die mogelijk voor het CBS in de (nabije) toekomst nieuwe bruikbare databronnen opleveren. Mogelijk nieuwe interessante databronnen zijn:

- Slimme elektriciteit- en gasmeters (gegevens over energieverbruik eventueel zelfs inclusief tips voor bezuinigen; http://www.utilities.nl/index.php?option=com_content&task=view&id=1765&Itemid=39).

⁴ Dit hoeft niet altijd het geval te zijn, denk bijvoorbeeld aan een website waarbij de prijzen van producten alleen als onderdeel van een afbeelding of een animatie beschikbaar zijn. Ook de prijzen die interactief uit een bedrijfsdatabase worden gehaald zijn via deze weg moeilijker te vinden. Een voorbeeld van een succesvolle toepassing zijn de prijsvergelijkingswebsites, zoals die van Kelkoo en vergelijk.nl.

- eFormulieren (elektronische formulieren van overheidsinstellingen; <http://www.e-overheid.nl/sites/eformulieren/>)
- Rekeningrijden (http://www.verkeerenwaterstaat.nl/onderwerpen/mobiliteit_en_bereikbaarheid/anders_betalen_voor_mobiliteit/)
- eFacturen (elektronische facturen; <http://www.platformfa.nl/>)
- Digitaal klantdossier (elektronische gegevens uit de keten werk en inkomen; <http://www.dkd.nl>)
- Geografische coördinaten van activiteiten (ter aanvulling van het ‘CBS in uw buurt’ initiatief; <http://www.cbsinuwbuurt.nl/>).

Ook banktransacties kunnen een interessante bron van informatie gaan vormen. De goede contacten met de Nederlandse Bank kunnen verder uitgebouwd worden om te bepalen waar hier de mogelijkheden liggen; bijvoorbeeld het gebruik van de betalingstransactie gegevens van de Society for Worldwide Interbank Financial Telecommunication (SWIFT).

In het algemeen geldt dat de gegevens van alle elektronica die op de één of andere manier met het internet verbonden zijn interessant kunnen zijn voor het CBS wanneer ze met sociale en economische activiteiten samenhangen.

3. Conclusie

In dit rapport zijn potentieel interessante databronnen voor het CBS besproken. Waar mogelijk zijn toepassingen genoemd. Doordat steeds meer apparaten de mogelijkheid bezitten om gegevens op te slaan en deze met andere apparaten en systemen uit te wisselen is te verwachten dat er voor het CBS steeds meer interessante nieuwe databronnen beschikbaar komen. Hoe beter het CBS van die nieuwe bronnen gebruik maakt, hoe meer cijfers ze over sociale en economische ontwikkelingen in Nederland kan produceren zonder daarmee personen en bedrijven te belasten. Uiteraard dient hierbij de waarborging van de privacy van de betrokkenen voorop te staan (CBS 2004). Voor het bepalen van de bruikbaarheid van de in dit rapport beschreven databronnen zal nader onderzoek dienen te worden verricht. Concrete plannen in deze richting zijn de volgende:

- Voor het gebruik van XBRL-GL voor de internationale handel gegevens is een voorstel in het kader van MEETS ingediend. MEETS (Modernisation of European Enterprise and Trade Statistics) is een Eurostat programma voor de modernisering van de Europese bedrijfs- en handelsstatistiek (MEETS 2007). Het resultaat van deze aanvraag is nog niet bekend.
- Bij de taakgroep Verkeer en Vervoer zal een AIO o.a. onderzoek gaan verrichten naar de bruikbaarheid van GPS-gegevens voor de verkeer- en vervoerstatistiek en naar RFID-gegevens voor de internationale handel (Signaal 2008).
- Het gebruik van het internet als (aanvullende) informatiebron is voorgesteld als terugvalscenario voor statistieken die van registergegevens afhankelijk zijn. Webpagina's van bedrijven worden ook nu reeds op het CBS gebruikt om ontbrekende gegevens snel op te zoeken. Daarnaast is er in het onderzoeksprogramma van Methodologie voor 2008 (DMK en O&O) opgenomen dat verder onderzoek zal worden verricht naar geautomatiseerde waarneming van internetgegevens via 'webspiders' (DMK 2008).

De bovenstaande lijst laat zien dat voor onderzoek naar de bruikbaarheid van potentiële bronnen reeds initiatieven zijn ontplooid op het CBS. Twee onderwerpen ontbreken nadrukkelijk. Het eerste is onderzoek naar het gebruik van GSM-gegevens. Het is mogelijk dat deze gegevens ook in het promotieonderzoek bij Verkeer en Vervoer bekeken gaan worden; b.v. door gebruik te maken van de GSM-verplaatsingsgegevens van de chauffeur. Het is in ieder geval aan te raden dit onderwerp in het onderzoeksprogramma van DMK voor 2009 op te nemen. Omdat er voor het verkrijgen van GSM-gegevens nog contacten moeten worden gelegd zal er dit jaar al een begin mee moeten worden gemaakt. Het tweede onderwerp dat nadrukkelijk ontbreekt is een (voor)onderzoek naar de bruikbaarheid van gegevens op internet, zoals die in weblogs, op sociale netwerksites en koop- en verkoopsites. Dit onderwerp dient zeker nader bekeken te worden, bij voorkeur in samenwerking

met een universiteit. Hierbij dient uiterst zorgvuldig om te worden gegaan met privacygevoelige gegevens.

Literatuur

Adams, C. (2005) How much of all internet traffic is pornography? The Straight Dope. (<http://www.straightdope.com/columns/051007.html>).

Betaalwijzer (2006) Nieuwsbrief over elektronisch betalen, nr. 1, juli (<http://www.chipknip.nl/Chipknip.nl/pdf/Betaalwijzer%202006%20Juli%20-%2012%20pag%20-%20web%20versie.pdf>).

Bethlehem, J. (2008) Reactie op rapport “Internet as a Data source”, 22-mei.

Blogresearch (2008) Toenemend belang van weblogs. Website van het Nederlandse marktonderzoeksbureau blogresearch.nl (<http://www.blogresearch.nl/?WEBLOG/979>).

Burger, J.M.S., Roos, M.R. (2009) Potential of XBRL dimensions for International Trade Statistics, Discussion paper. (<http://www.cbs.nl/>

CAIDA (2008) Website van de Cooperative Association for Internet Data Analysis. (<http://www.caida.org/home/>).

CBS (2004) Wet op het Centraal Bureau voor de Statistiek. Staatsblad 695, 15 december.

CBS (2008) De digitale economie 2007, inclusief statistische bijlage. (<http://www.cbs.nl/nl-NL/menu/themas/bedrijven/publicaties/digitale-economie/publicaties/2008-de-digitale-economie-p34-pub.htm>).

Daas, P.J.H., Roos, M. (2008) Benaderingsstrategieën, Deelthema: XBRL. Deelthema in de reeks van statistische methoden, CBS Heerlen/Voorburg, 6 maart.

DGOF (2008) Website van de ‘Deutschen Gesellschaft für Online-Forschung’ (<http://www.dgof.de/>)

Dialogic, Ministerie van Economische Zaken, Univ. Utrecht (2008) Go with the dataflow!, Analysing the Internet as a data source (IaD). Rapport inclusief bijlage, 13-mei (http://www.ez.nl/Onderwerpen/Elektronische_communicatie/ICT_beleid/Consultatie_rapport_Go_with_the_Dataflow_Analysing_the_Internet_as_a_data_source_IaD).

DMK (2008) Onderzoeksprogramma CBS 2008. Versie 071130, d.d. 20 oktober.

Gautier, P., Siegmann, A., van Vuuren, A. (2006) Het effect van de moord op Theo van Gogh op huizenprijzen in Amsterdam. Tijdschrift voor Politieke Economie, jaargang 27, vol. 6, pp. 5-25 (http://www.tinbergen.nl/~siegmann/download/TPE_Theo.pdf).

GPX (2008) GPX: the GPS exchange format (<http://www.topografix.com/gpx.asp>).

Hoffman, C. (2006), Financial Reporting Using XBRL: IFRS and US GAAP Edition, UBMatrix Inc. Lulu Enterprises, Sevilla, Spain.

Homburg, G.H.J., Dekkers, S. (2003) Cameratoezicht in de openbare ruimte. Rapport 1, College bescherming persoonsgegevens, Den Haag, november (http://www.cbppweb.nl/downloads_rapporten/rap_2003_cameratoezicht.pdf?refer=true&theme=purple).

InstatXML (2003) Overzicht van het Instat/XML project, (<http://www.ebxml.eu.org/Dutch/INSTAT.htm>).

Internet statistics guide (2002) Complete Guide to Internet Statistics and Research. (<http://internet-statistics-guide.netfirms.com/>).

Internetresearchers (2008) Website of the Association of Internet Researchers. (<http://aoir.org>).

Ipoque (2006) P2P Survey 2006 (http://www.ipoque.com/userfiles/file/p2p_survey_2006.pdf) .

Ipoque (2007) Internet Study 2007 (http://www.ipoque.com/userfiles/file/Internet_study_2007_abstract_en.pdf) .

Jones, S. (1998) Doing Internet Research: Critical Issues and Methods for Examining the Net. Sage Publications, Inc. California, USA.

Koning, B., de (2008) Alles onder controle, De overheid houdt u in de gaten. Uitgeverij Balans, Amsterdam.

MEETS (2007) Programme for the Modernisation of European Enterprise and Trade Statistics MEETS (<http://www.europarl.europa.eu/oeil/OpenDetailFiche.do?ficheId=1184&language=en>).

Mijnprivacy (2008) Informatieve website van het College Bescherming Persoonsgegevens (<http://www.mijnprivacy.nl/Home.htm>).

Monk, T., Claffy, K.C. (1996) A Survey of Internet Statistics / Metrics Activities. 18 juni (<http://www.caida.org/publications/papers/1996/metricsurvey/metricsurvey.html>).

Moreno, Y., Pastor-Satorras, R., Vespignani, A. (2002) Epidemic outbreaks in complex heterogeneous networks. Eur. Phys. J. B 26, 521-529.

Muller, H., Buijs, C., Timmermans, G. (2007) Administratieve lasten CBS - 2006 (definitieve cijfers). Divisie Bedrijfseconomische Statistieken, Centraal Bureau voor de Statistiek, 24 september.

Pasmooij, J., Haas, M., de (2005) Nederlands Taxonomie Project maakt aanlevering financiële gegevens makkelijker en goedkoper. Accountancynieuws 11, pp. 12-14. 3 juni (<http://www.xbrl-ntp.nl/publicaties/050602104144.pdf>).

Signaal (2008) CBS en RSM Erasmus Universiteit werken aan verbetering handels- en transportstatistieken. Signaal, Januari, nr. 1, p. 9.

Snijkers, G, Göttgens, R., Luppens, M. (2003) WaarnemingsExpertise Programma(WEP): Naar een professionele waarneming. Centraal Bureau voor de Statistiek Heerlen.

Tan, A-H. (1999) Text mining: The state of the art and the challenges. Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining, workshop on Knowledge Discovery from Advanced Databases, pp. 65-70, (http://citeseer.ist.psu.edu/cache/papers/cs/12174/http:zSzzSztextmining.krdl.org.sgzSzdscszSztext_mining_KDAD_99.pdf/tan99text.pdf)

Technorati (2008) About us page. Website van de weblog autoriteit technorati (<http://technorati.com/about/>).

Telecomwereld (2008) Dienst Waarbenik spoort personen op via mobieltje (<http://www.telecomwereld.nl/n0000278.htm>).

Tweede Kamer (2007) Wijziging Telecommunicatiewet i.v.m. Richtlijn dataretentie, Memorie van toelichting. Vergaderjaar 2006-2007, publicatienummer 31145, nr. 3, Den Haag.

Tweede Kamer (2008) Beslissing wijziging Telecommunicatiewet i.v.m. Richtlijn dataretentie, Memorie van toelichting. Vergaderjaar 2006-2007, publicatienummer 31145, nr. 3, Den Haag.

Webwereld (2007) TomTom baseert filegegevens op lokatie Vodafone-abonnees (<http://webwereld.nl/articles/48647/muslix64>).

WWE (2006) Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Website of the 3rd Annual workshop on Weblogging, Edinburgh (<http://www.blogpulse.com/www2006-workshop/>).

XBRL (2006) XBRL Dimension 1.0, Recommendation, 18 september. (<http://www.xbrl.org/Specification/XDT-REC-2006-09-18.htm>).

XBRL (2008) Global Ledger Taxonomy - An Introduction. XBRL.org website (<http://www.xbrl.org/GLTaxonomy/>).

Bijlage 1: Bruikbaarheid nieuwe bronnen voor de verschillende CBS-thema's

Thema's	Nieuwe bronnen	Bedrijfs- administraties	RFID	GPS	GSM	Internet
Arbeid en sociale zekerheid		+				+/-
Bedrijven		++				+
Bevolking						+/-
Bouwen en wonen		+/-				++
Financiële en zakelijke diensten		++				+/-
Gezondheid en welzijn					+/-	+/-
Handel en horeca		++	+			+/-
Industrie en energie		++	+			+/-
Inkomen en bestedingen					+/-	+
Internationale handel		++	+	+	+/-	+/-
Landbouw		+				+/-
Macro-economie						
Natuur en milieu				+	+	+/-
Onderwijs						+
Overheid en politiek						+/-
Prijzen		++				+
Veiligheid en recht						+/-
Verkeer en vervoer		++	+	++	+	+/-
Vrije tijd en cultuur				+/-	+	+
Dossiers						+

Verklaring van de tekens: ++, zeer bruikbaar; +, bruikbaar, +/-, mogelijk bruikbaar. Omdat het internet veel gegevens bevat zou deze bron -in principe- voor veel CBS-thema's (aanvullende) informatie kunnen opleveren. Momenteel is echter niet duidelijk hoe bruikbaar deze informatie is; vandaar de vele +/- in deze kolom. Hiervoor is aanvullend onderzoek noodzakelijk.

Bijlage 2: Afkortingenlijst

ADSL	Asymmetric Digital Subscriber Line
AIO	Assistent In Opleiding
Blog	Weblog
CAIDA	Cooperative Association for Internet Data Analysis
CBS	Centraal Bureau voor de Statistiek
CPI	Consumenten Prijs Index
GL	Global Ledger
CO ₂	Koolstofdioxide
DMK	Divisie Methodologie en Kwaliteit
FTP	File Transfer Protocol
GPS	Global Positioning System
GPX	Global Positioning System eXchange formaat
GSM	Global System for Mobile communications
GUID	Globally Unique Identifier
ICQ	Engelstalige fonetisch afkorting, klinkt uitgesproken als: I seek you
IMEI	International Mobile Equipment Identity
IP	Internet Protocol
IPv4	Internet Protocol versie 4
IPv6	Internet Protocol versie 6
MEETS	Modernisation of European Enterprise and Trade Statistics
MSN	Microsoft Network
O&O	Ontwikkeling en Ondersteuning
P2P	Peer-to-Peer
QoS	Quality of Service
RFC	Request For Comments
RFID	Radio frequency identification
SIM	Subscriber Identity Module
SWIFT	Society for Worldwide Interbank Financial Telecommunication
TCP	Transmission Control Protocol
TOS	Type Of Service
TTL	Time To Live
UDP	User Datagram Protocol
UPC	United Pan-Europe Communications
XBRL	eXtensible Business Reporting Language
XBRL-GL	eXtensible Business Reporting Language Global Ledger
XML	eXtensible Markup Language

Bijlage 3: Internet protocol in detail

Iedere computer (beter: elke netwerkkaart) die via het internet met andere computers communiceert moet een uniek IP-adres bezitten. Een IP-adres bestaat uit een combinatie van aantal sets van één tot drie cijfers gescheiden door punten. Een voorbeeld van een correct IP-adres is 123.233.255.255. Het laagste nummer is 0 en het hoogste nummer is 255. Voor specifiek gebruik zijn bepaalde sets van IP-adressen gereserveerd. Zo zijn de IP-adressen die beginnen met 10.0.0.0/8, 172.16.0.0/12 en 192.168.0.0/24 gereserveerd voor lokale netwerken.

Het toekennen van IP-adressen aan computers (of andere apparaten) die direct met het internet verbonden zijn gebeurt door de internetproviders. Deze kan de IP-adressen statisch of dynamisch toekennen. Internetproviders die statische adressen gebruiken zorgen ervoor dat de computer (of een ander apparaat zoals een router) van een gebruiker altijd hetzelfde IP-adres toebedeeld krijgt. Dit maakt een computer met een statisch IP-adres uniek identificeerbaar. Bij gebruik van dynamische IP-adressen kan het IP-adres dat aan een computer (of een router) wordt toegekend variëren. Echter, doordat elke internetprovider een beperkte set van IP-adressen gebruikt zullen vaak alleen de laatste drie cijfers variëren. Het toegekende IP-adres blijft in gebruik totdat de computer wordt uitgezet of wanneer deze opnieuw wordt opgestart.

Door het grote succes van het internet, worden steeds meer computers op het internet aangesloten. Dit bracht een beperking in het tot dan toe gebruikte Internet Protocol aan het licht. Met die versie, versie 4 (IPv4) zijn er grofweg 4,2 miljard unieke IP-adressen te genereren. Dat lijkt veel, maar een groot deel is al gereserveerd voor allerlei subnetwerken, die daar weer gedeelten van gebruiken. Bovendien mag een groot aantal adressen niet gebruikt worden omdat die voor specifiek gebruik gereserveerd zijn (b.v. de adressen voor lokale netwerken). Dit heeft geleid tot het ontwikkelen van een nieuwe IP-versie (versie 6) waarmee een vrijwel oneindige hoeveelheid unieke adressen gebruikt kunnen worden. Momenteel wordt IPv6 nog vrijwel uitsluitend in Japan en Korea gebruikt. Voordeel van IPv4 is dat dit protocol door alle computers op het internet wordt ondersteund.

IPv4

Een IPv4-adres bestaat uit 32 bits, meestal weergegeven als vier 8-bits getallen, gescheiden door punten, waarbij ieder getal de waarde 0 t/m 255 kan bevatten. Het ene deel van deze 32 bits geeft het netwerknummer aan, het andere deel het hostnummer. Een apart netmasker bepaalt welke bits bij het netwerknummer horen, en welke bij het hostnummer. Hosts (computers) met hetzelfde netwerknummer kunnen direct met elkaar communiceren, hosts met een verschillend netwerknummer kunnen dat niet, en maken gebruik van een of meer tussenliggende routers om hun doel te bereiken.

Een IPv4-pakket moet de volgende onderdelen hebben. De hoeveelheid ruimte die voor ieder onderdeel gereserveerd is, is tussen haakjes gegeven.

De zogenaamde headers:

- Versie (4 bits). In dit geval 4.
- Headerlengte (4 bits). Lengte van de header, uitgedrukt in woorden van 32 bits. De minimum lengte is 5 woorden. Als dit veld bijvoorbeeld 0111 is, is de grootte van de header 28 bytes (0111 is 7 in decimale waarde, $7 * 32 = 224 \text{ bits} = 28 \text{ bytes}$).
- Differentiated Service/Type Of Service (1 byte). Dit speelt een rol bij het bepalen van de route die het pakket aflegt. Het wordt tegenwoordig gebruikt voor Quality of Service (QoS). Het TOS-veld (4 bits) is sinds RFC 791 als volgt gedefinieerd:
 - o 0000 Standaard
 - o 0001 Minimale "kosten": lage prioriteit
 - o 0010 Maximale betrouwbaarheid
 - o 0100 Maximale doorvoersnelheid
 - o 1000 Minimale wachttijd
 - o 1111 Maximale beveiliging
- Totale lengte in bytes (2 bytes): de grootte van het pakket (gebruikt om vast te stellen of er iets verloren is gegaan en om te weten waar het pakket ophoudt).
- Identificatie (2 bytes). Nodig voor het uitfilteren van foute pakketten en het aan elkaar plakken van gefragmenteerde data.
- Vlaggen (3 bits). De eerste vlag is (nog) ongebruikt, hij moet 0 zijn. Verder: mag dit pakket gefragmenteerd worden? Is dit pakket een fragment?
- Fragmentbegin (13 bits). Waar het fragment begint.
- Time To Live - TTL (1 byte). Aantal hops waarna het pakket moet verdwijnen. Iedere hop verlaagt de TTL van een pakket met 1. Als TTL 0 wordt, zal het pakket niet verder gestuurd worden.
- Protocol (1 byte): het protocol dat binnen dit IP-pakket gebruikt wordt, zoals UDP of TCP.
- Controlesom van headers (2 bytes). Gebruikt om corrupte pakketten te filteren.
- Bron (4 bytes): waar het IP-pakket vandaan komt.
- Bestemming (4 bytes). Waar het pakket naartoe moet.

Na de headers volgt het datagedeelte, dat bijvoorbeeld TCP of UDP bevat.

IPv6

Onder IPv6 zijn een aantal zaken anders, waaronder de grootte van de bron- en bestemmingsadressen. Deze bestaan uit zestien bytes. Dat is ook een reden waarom IPv6 steeds meer in zwang komt: IPv6 biedt een vrijwel oneindig aantal mogelijke adressen. Op 26 oktober 2007 is er door de Réseaux IP Européens een resolutie uitgevaardigd om de versnelde invoering van IPv6 te stimuleren. De Réseaux IP Européens is een samenwerkingsverband van personen en organisaties met een gezamenlijke interesse in IP-netwerken en het operatief houden van het internet.