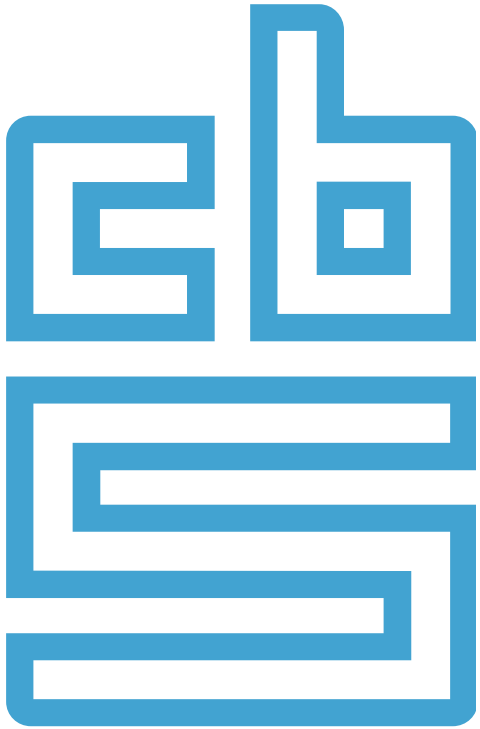


Uitdagingen bij datakoppelingen voor statistisch en surveyonderzoek

Presentatie voor het NPSO
11 februari 2019, Brussel

Eric Schulte Nordholt (e.schultenordholt@cbs.nl)
Centraal Bureau voor de Statistiek, Divisie Sociaal-economische en ruimtelijke statistieken

Inhoud

- Dataverzameling
- Voorbeelden van registers
- Definitie en redenen voor het SSB
- Koppelnummers
- Voorbeeld Volkstelling
- Koppelproces
- Conclusies
- Statistische beveiligingsaspecten
- Discussie

Dataverzameling

- Gecentraliseerd in één divisie dataverzameling (registers en surveys): efficiënter en professioneler
- Zo veel mogelijk gebruik maken van dezelfde infrastructuur voor sociale en economische statistieken
- Verzamelstrategie (voorkeursvolgorde):
  - registerdata
  - surveys
- Voor vele statistieken: alleen gebruik maken van bestaande bronnen

Voorbeelden van registers (1)

Drie soorten registers:

- Bevolkingsregister (BRP)
- Banenbestand
- Zelfstandigenbestand
- Opleidingsniveaubestand
- Beroepenregister
- Inkomensregister
- Register met gegevens over de sociale zekerheid

Voorbeelden van registers (2)

- Register met gegevens over werkloosheid
- Register met gegevens over pensioenen
- Andere registers over personen, families en huishoudens
- Woningregister
- Andere registers over percelen, gebouwen en woningen
- Bedrijfsregister (ABR)
- Andere registers over bedrijven en vestigingen

Definitie en redenen voor het SSB (1)

SSB: Systeem van sociaal-statistische bestanden

Definitie:
Set van geïntegreerde microdatabestanden met coherente en gedetaileerde demographische en sociaal-economische data over personen, huishoudens, banen en uitkeringen

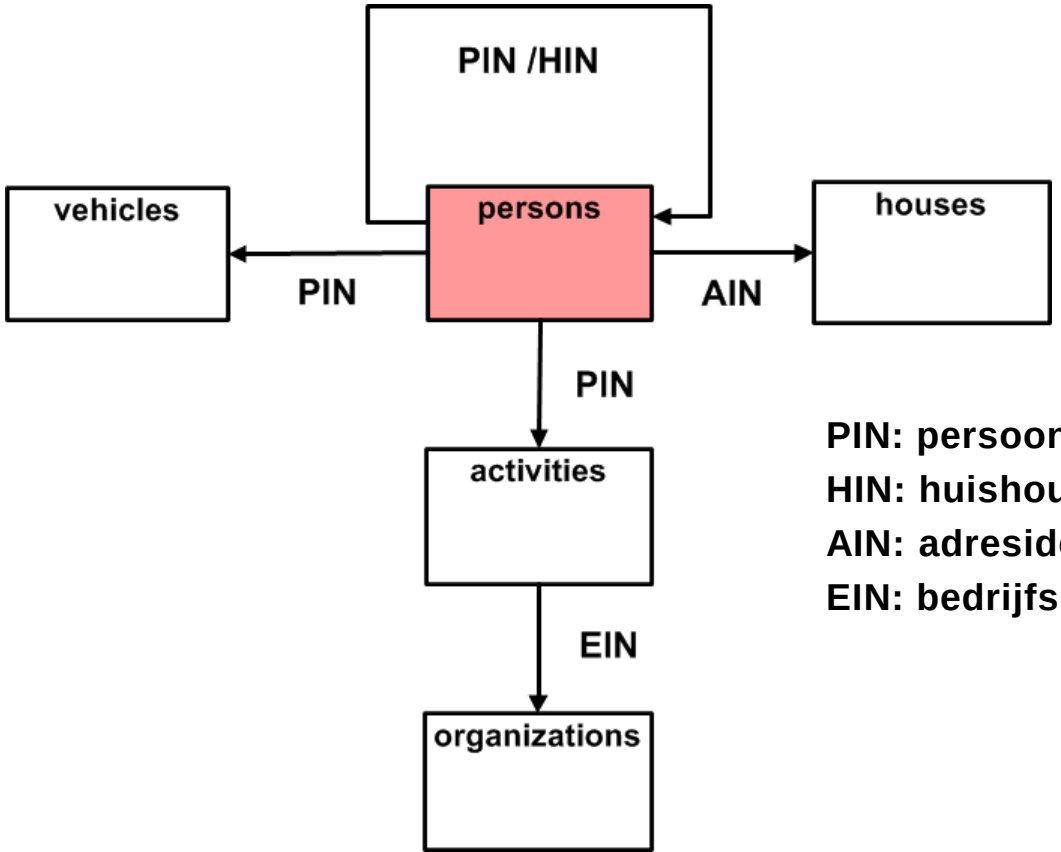
Geen interne inconsistente informatie

Definitie en redenen voor het SSB (2)

Redenen:

- Virtuele Volkstelling 2001
- Betere output: coherentie en flexibiliteit

Koppelnummers



PIN: persoonsidentificatienummer
HIN: huishoudensidentificatienummer
AIN: adresidentificatienummer
EIN: bedrijfsidentificatienummer

Voorbeeld Volkstelling

Bronnen

Voor de Volkstelling 2011:

Registers

- Bevolkingsregister (→ illegalen uitgesloten, daklozen op laatst bekende adres geteld)
- Banenbestand (alle werknemers)
- Zelfstandigenbestand (alle zelfstandigen)
- Informatie van de Belastingdienst
- Gegevens over sociale zekerheid
- Gegevens over pensioenen en levensverzekeringen
- Woningregister

Survey

- Enquête Beroepsbevolking (twee variabelen uit de EBB: beroep en opleidingsniveau)

Kenmerken Nederlandse Volkstelling (1)

- Louter gebruik maken van bestaande bronnen (registers en EBB) → afhankelijk van registerhouders (beschikbaarheid, tijdigheid)
- Kosten Nederlandse VT relatief gering (het CBS mag kosteloos gebruik maken van alle voor de officiële statistiek relevante registers) t.o.v. andere landen
- Late start: zodra alle relevante registers en enquêtes beschikbaar zijn

Kenmerken Nederlandse Volkstelling (2)

- Relatief korte doorlooptijd: geen grote nonrespons- en verwerkingsproblemen
- Consistente schattingen, maar niet alle details beschikbaar
- Aanvaardbaar alternatief voor de traditionele volkstelling
- Jaarlijkse updates mogelijk
- Vergelijkingen in de tijd en met andere landen

Koppelproces

Koppelproces (1)

- Koppelen van registers en andere datasets aan een zelf geconstrueerd Centraal Koppelbestand
- Records worden geïdentificeerd door een surrogaat identificatienummer (RIN)
- Een unieke tabel RIN-Burger Service Number (BSN)
- Minimale set van identificerende variabelen
- Elke stap in het proces is een deterministische koppeling

Koppelproces (2)

Centraal Koppelbestand **> 17 miljoen unieke personen**	
BSN	< 0,03% onbekend
Geboortedatum	< 0,5% onbekend (maand of dag)
Geslacht	Altijd
Postcode	< 0,05% onbekend
Huisnummer	< 0,05% onbekend
RIN Persoon	Altijd
RIN Adres	Altijd
Wanneer geldig?	Altijd

Koppelproces (3)

1. Koppelen op BSN
   Controleer geboortedatum en geslacht
   Een geldige koppeling als niet meer dan één van de variabelen geboortejaar, -maand, -dag en geslacht verschillen

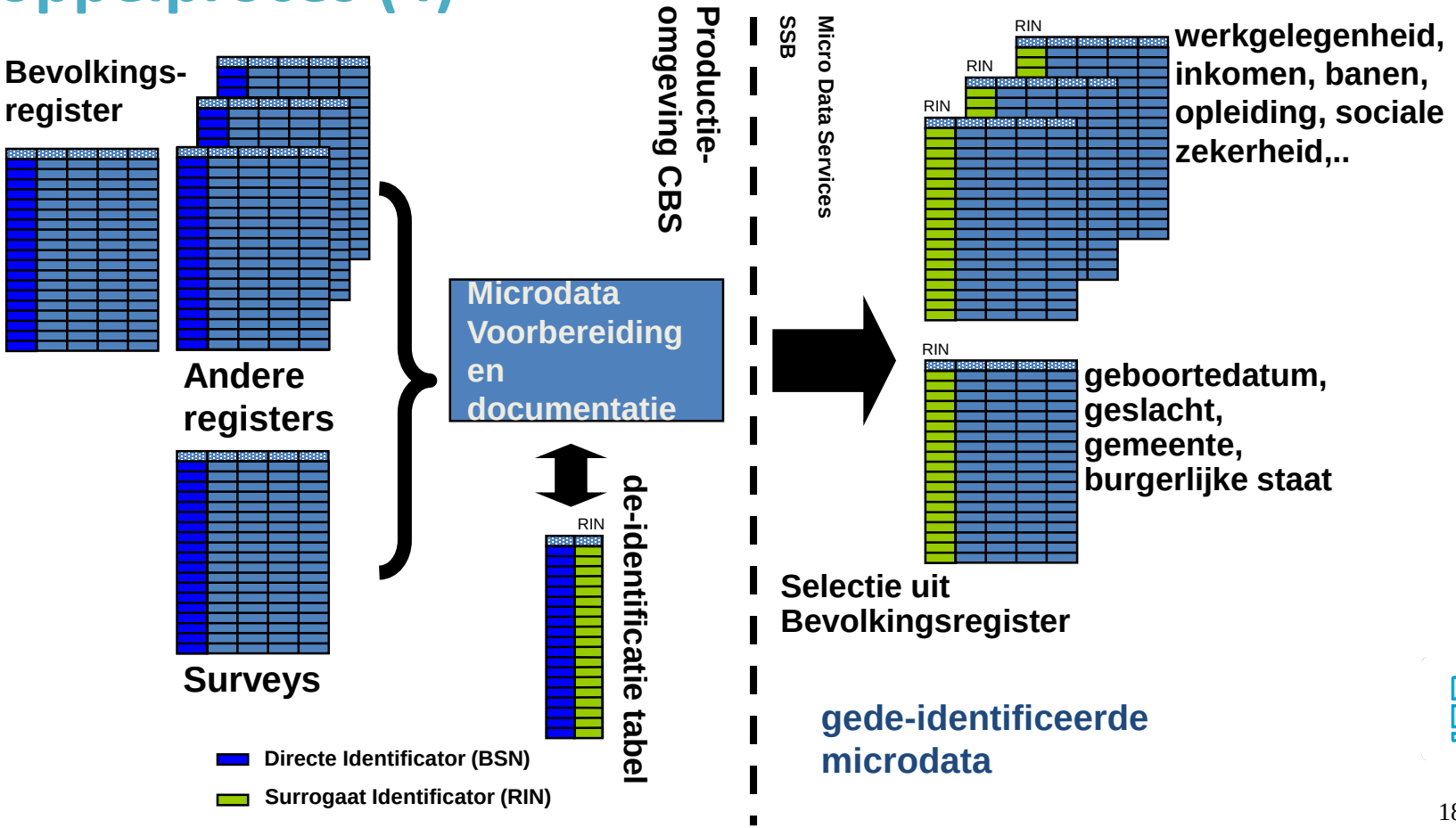
anders

2. Koppel op andere variabelen zoals postcode, huisnummer, geboortedatum, geslacht (alle koppelvariabelen moeten gelijk zijn)

anders

3. Koppel op BSN zonder controle op andere variabelen

Koppelproces (4)



Koppelproces (5)

Voorbeeld:

Statistiek Werkgelegenheid en Lonen		**3801246**	**100,0**
Totaal gekoppeld		**3747976**	**98,6**
1	BSN, geboortejaar, -maand, -dag, geslacht	3577090	94,1
2	Postcode, geboortejaar, -maand, -dag, geslacht	164267	4,3
3	BSN	6619	0,2
Niet gekoppeld		**53270**	**1,4**
	Geldig BSN	21194	0,6
	Geldige postcode	5799	0,2
	Ongeldige postcode	10294	0,3
	Niet-inwoner	5101	0,1
	Onbekend of ongeldig BSN	**32076**	**0,8**
	Geldige postcode	8718	0,2
	Ongeldige postcode	20052	0,5
	Niet-inwoner	3306	0,1

Conclusies

- Koppelen is relatief goedkoop
- Koppelen is relatief snel (korte productietijd)
- Het SSB heeft zijn plaats in de organisatie gevonden
- Statistische beveiligingsaspecten zijn veel belangrijker geworden!

Statistische beveiligingsaspecten (1)

Algemene Verordening Gegevensbescherming (AVG)

CBS moet er op grond van de AVG voor zorgen dat output geen identificeerbare persoonsgegevens bevat, maar de AVG geeft niet aan hoe dat moet worden bereikt

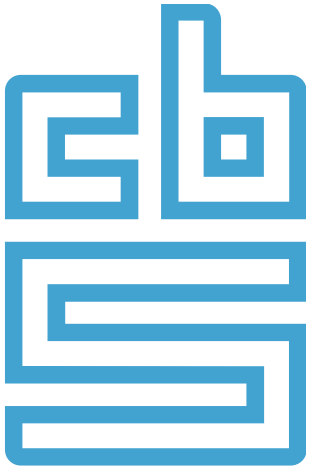
CBS-wet: uit output mogen geen herkenbare gegevens over individuele personen kunnen worden afgeleid

Statistische beveiligingsaspecten (2)

- Privacy altijd gewaarborgd (in de wet vastgelegd)
- Informatiebeveiliging (gebouw en IT)
- Geheimhoudingsplicht voor alle medewerkers
- Gegevens over mensen worden direct gescheiden van de namen en de adressen
- De wet garandeert dat gegevens alleen voor statistische doeleinden worden gebruikt
- Microdata voor onderzoek (eventueel na pseudonimisering door een TTP) zijn niet open maar wel FAIR (Findable, Accessible, Interoperable en Reusable)

Discussie

Zijn er vragen of opmerkingen?



Voor wat er **feitelijk** gebeurt