Hoofdlijnen van de Business Architectuur CBS

Robbert Renssen

Publicatiedatum CBS-website: 29 juni 2010

Centraal Bureau voor de Statistiek

Centraal Bureau voor de Statistiek

Verklaring van tekens

.	=	gegevens ontbreken
*	=	voorlopig cijfer
**	=	nader 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
2008–2009	=	2008 tot en met 2009
2008/2009	=	het gemiddelde over de jaren 2008 tot en met 2009
2008/'09	=	oogstjaar, boekjaar, schooljaar enz., beginnend in 2008 en eindigend in 2009
2006/'07–2008/'09	=	oogstjaar, boekjaar enz., 2006/'07 tot en met 2008/'09

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 - Grafimedia

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

Hoofdlijnen van de Business Architectuur CBS

Robbert Renssen[1]

Samenvatting: in dit cursusdocument worden de hoofdlijnen van de business architectuur van het CBS uiteengezet. Het niveau van deze inleidende cursus is 'intermediate'. De nadruk ligt niet zozeer op allerlei theoretische begrippen uit de business architectuur, maar op de toepassing ervan op het CBS. Het doel van de cursus is dat de cursist een idee krijgt welke aspecten van belang zijn bij het 'maken van een statistiek', en hoe deze aspecten in samenhang tot elkaar staan. En dat is ook precies wat architectuur beoogt: het geven van een samenvattend beeld van één of meer artefacten in een omgeving, vooral betreffend coherente keuzes op het gebied van functie, structuur en stijl. De twee belangrijkste onderwerpen die in de cursus aan de orde komen zijn 1) het statistische product, i.e. wat is een statistiek en 2) het statistische proces, i.e. hoe maken we een statistiek.

De cursus biedt een leidraad voor managers, projectleiders, projectmedewerkers, methodologen, ict-ers, statistici, en anderen, die te maken gaan krijgen met herontwerpen van statistische processen.

Trefwoorden: architectuurprincipes, attribuutvariabele, BAD, benadereenheid, businessdomein, classificatiebase, classificerende variabele, conceptuele metadata, eenhedenbase, fysiek datamodel, geldigheidsperiode, GSBPM, (her)ontwerp, IAF, ketenregie, klasse, koppelvlak, kwaliteitsmetadata, kwantificerende variabele, logisch datamodel methodologie, niet-inhoudelijke (technische) transformatie, niet-reële mutatie, objecttype, populatie, populatiekader, primaire waarneming, procesdienst, procesmetadata, procesregie, reële mutatie, regelsturing, ruggengraat, rustpunt, secundaire waarneming, sleutelvariabele, standaard processtap, statistisch proces, statistisch product, waarneemeenheid.

---

[1] De visie in dit document is de visie van de auteur en niet noodzakelijkerwijs CBS-beleid. Bij het ontwikkelen van deze visie heb ik dankbaar gebruik gemaakt van de vele discussies met mijn CBS-collega's. Zij hebben mij daarmee, veelal onbewust, geholpen met de uiteindelijke totstandkoming van dit document.

Woord vooraf

Dit document is een tweede herziene versie van de hoofdlijnen van de Business Architectuur CBS uit 2008 (BPA nr: DMK-2008-11-07-RRNN). Ten opzichte van de eerste versie is de Business Architectuur op een aantal onderwerpen verder uitgewerkt en zijn een aantal onderwerpen toegevoegd.

Zo wordt dieper in gegaan op het waarnemen van registers aan de inputkant en is nu ook de outputkant in beeld gebracht. Tevens zijn de concepten keten(regie) en standaard processtappen aan de verwerkingskant verder uitgewerkt.

Ten slotte heeft deze druk een aantal redactionele wijzigingen ondergaan, die hopelijk de begrijpelijkheid en de leesbaarheid van de stof zullen vergroten.

Robbert Renssen

Inhoudsopgave

1. Inleiding

Het CBS ziet zich de komende jaren voor verschillende uitdagingen geplaatst, aldus de eerste nieuwsbrief van het vernieuwingsprogramma dat in jaar 2007 van start ging. Zo staan verhoging van de efficiency en verlaging van de administratieve lastendruk voor het bedrijfsleven hoog op de politieke agenda.

Door het herontwerpen van statistische processen volgens voorgeschreven architectuur en met gebruikmaking van generieke procesdiensten dient een toekomstbestendig CBS te worden ingericht. Kenmerkend van dit CBS is standaardisatie, bijvoorbeeld op het gebied van data-uitwisseling, van waarneming, van verwerking, van statistische methoden, etc.

De uitdaging is om *tijdens* een herontwerp met de genoemde standaarden rekening te houden, niet *achteraf*. Zolang de gekozen standaarden onderling niet strijdig zijn en het grootste deel van de statistiekproductie in de gekozen standaardisatie past zijn de keuzes goed. Als dit zonder architectuur kan, prima. Maar de ervaring leert dat statistiek maken een complex proces is met veel historisch gegroeide oplossingen. Langzamerhand is het overzicht kwijtgeraakt. Enige vorm van structurering om het overzicht weer terug te krijgen en zodoende de juiste standaarden te kiezen lijkt onvermijdelijk.

Wat is nu architectuur? Er zijn vele definities, maar volgens Wagner is het een *samenhangend geheel* van principes en modellen dat *richting* en *structuur* geeft in ontwerp en realisatie van de processen, organisatorische inrichting, informatie-voorziening en technische infrastructuur.

Wat is de toegevoegde waarde van architectuur? Het reduceert de complexiteit van statistische processen door het aanbrengen van structuur. Het beschrijft deze processen op uniforme wijze in een taal en in symbolen die algemeen en neutraal voor communicatie toepasbaar zijn. Het biedt ondersteuning bij het maken van keuzes binnen en over herontwerpen. Het is een middel om bij een herontwerp de samenhang tussen (IT-) oplossingsrichtingen zichtbaar te maken. Het geeft een totaalschets van de toekomstige situatie welke als richtpunt voor ontwikkeling kan dienen.

Deze inleidende cursus over architectuur beperkt zich tot de business architectuur van het CBS met betrekking tot het ‘ideale’ statistische proces. Er komen zes onderwerpen aan bod. Het eerste onderwerp (deel A) betreft een Integrated Architectural Framework (IAF). Dit framework helpt om de vele verschillende invalshoeken over architectuur te kunnen positioneren. Het tweede en derde onderwerp gaat over het ***product*** dat het CBS produceert (deel B) en het ‘ideale’ ***statistische proces*** dat daarvoor is ingericht (deel C). Het vierde onderwerp (deel D) legt de Business Architectuur van het CBS uit aan de hand van deel B en C. Het vijfde onderwerp (deel E) gaat in op (her)ontwerpen van statistische processen onder architectuur. Het zesde en laatste onderwerp (deel F) legt de relatie met een aantal verwante architecturen en gaat kort in op een aantal algemene modelleringstalen waarmee onderdelen van architectuur kunnen worden vastgelegd.

Deel A: Integrated Architectural Framework

2. Integrated Architectural Framework (IAF)

Eén van de manieren waarop architectuur complexiteit reduceert is door een organisatie, zoals het CBS met haar statistische processen, vanuit verschillende invalshoeken te bezien en deze invalshoeken zoveel mogelijk uit elkaar proberen te houden. Daartoe heeft het CBS het zogenaamde Integrated Architectural Framework (IAF) geadopteerd. In dit framework worden vier aspectgebieden onderkend:

- Business: dit aspectgebied beschrijft de (reële) producten, processen, actoren, middelen, doelen, organisatieonderdelen, etc. van een bedrijf
- Informatie: dit aspectgebied beschrijft de informatiekant van de (reële) producten, processen, actoren, middelen, doelen, organisatieonderdelen, etc van een bedrijf[2]
- InformatieSystemen: dit aspectgebied beschrijft de 'softwarematige kant' die nodig is voor de geautomatiseerde informatie
- Technische Infrastructuur: dit aspect gebied beschrijft de 'hardwarematige kant' die nodig is voor de 'softwarematige kant'.

Bij het inrichten van een statistisch proces zal aan alle bovengenoemde aspectgebieden de nodige aandacht moeten worden gegeven. Daarin zit een bepaalde volgorde in. Om een technische infrastructuur te kunnen inrichten, dient bekend te zijn welke 'software' nodig is. Om de software te kunnen bepalen moet duidelijk zijn wat de (geautomatiseerde) informatiebehoefte is en om dat in kaart te brengen moeten de businessdoelen, etc. helder zijn.

Indien het inrichten van een statistisch proces wordt vertaald naar het oplossen van één of meer problemen op ieder van de aspectgebieden, dan kunnen voor al deze aspectgebieden nog de volgende lagen worden onderscheiden:

- Context: deze laag beschrijft de aanleiding van het probleem en de omgevingsfactoren waarmee de oplossing van het probleem mee te maken heeft.
- Concept: deze laag beschrijft het probleem zelf, zonder dat op de oplossing wordt ingegaan.
- Logical: deze laag beschrijft de ideale (theoretische) oplossing, d.w.z. de ideale inrichting van een statistiek, waarbij nog geen specifieke keuzes zijn gemaakt t.a.v. de daadwerkelijke realisatie.

---

[2] De aspectgebieden business en informatie zijn, althans voor de auteur, soms moeilijk te onderscheiden omdat het statistische product van het CBS ook informatie is, maar voor de Nederlandse samenleving.

– Physical: deze laag beschrijft de daadwerkelijke realisatie richting ideale oplossing, waarbij eventueel meerdere mijlpalen worden onderscheiden. Pas in deze laag worden keuzes in bijvoorbeeld softwaretools gemaakt.

In figuur A-1 is het Integrated Architectural Framework (IAF) uitgebeeld, waarbij de aspectgebieden verticaal zijn weergegeven en de lagen daarin horizontaal. De belangrijkste meerwaarde van het IAF is niet zozeer de indeling zelf – er zijn meerdere frameworks met net iets andere indelingen – als meer het ***besef*** dat vanuit meerdere aspectgebieden op verschillende abstractieniveaus hetzelfde statistische proces kan worden bezien.

Figuur A-1: IAF

De onderwerpen van de resterende delen van deze cursus concentreren zich voor een groot deel op de afgevinkte cellen.

Deel B: Het Statistische Product

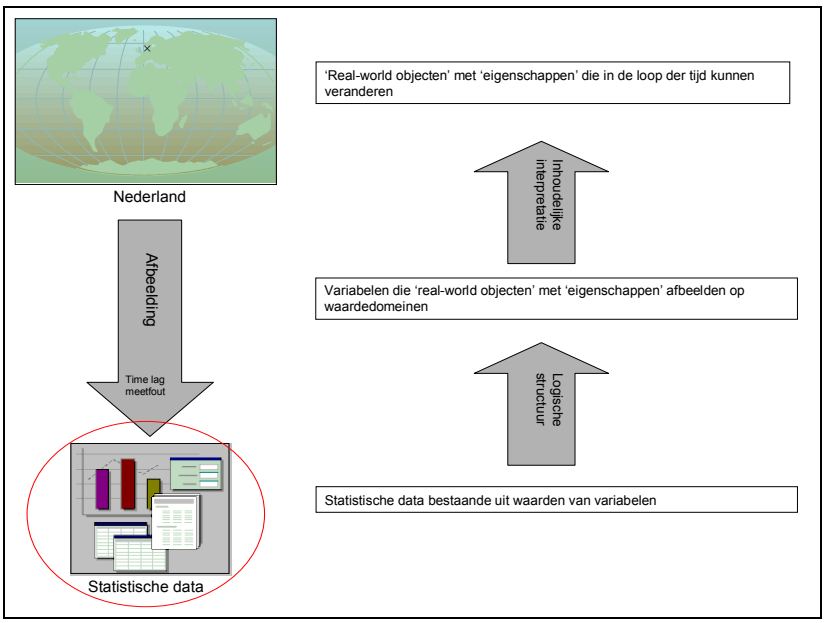
De missie van het Centraal Bureau voor de Statistiek (CBS) is het publiceren van betrouwbare en samenhangende statistische informatie, die inspeelt op de behoefte van de samenleving. Het CBS houdt zich daarmee voor een groot deel bezig met beschrijvend statistisch onderzoek: zij bepaalt van een deel van een populatie de waarden van variabelen en bewerkt deze tot descriptieve maten – statistische data – voor de gehele populatie of delen hieruit.

In een notendop is hiermee gezegd wat het belangrijkste product van het CBS is en hoe het CBS dit product maakt. In het tweede deel van deze inleidende cursus zal op het product worden ingegaan.

3. Contouren van het statistische product

Het statistische product van het CBS bevat statistische data, i.e. data die eigenschappen van real-world objecten afbeelden, ten behoeve van de informatievoorziening voor de samenleving. Als het CBS een autobedrijf was geweest dan zou het CBS geen statistische data als product hebben, maar auto's of onderdelen van auto's. In figuur B-1 is het statistische product als een afbeelding schematisch weergegeven.

Figuur B-1: Feiten in Nederland afgebeeld op een dataset



Ten aanzien van dit product kan een aantal aspecten worden onderscheiden, namelijk (een keuze van) de afbeelding, een inhoudelijke beschrijving van de afbeelding, een logische structuur van de afbeelding en een fysieke dataset (met een technische structuur met keuzes voor formaat, tekenset en syntax).

3.1 Keuze en beschrijving van een afbeelding

3.1.1 Conceptuele metadata

Bij de inhoudelijke keuze van een afbeelding hanteren we de volgende begrippen en uitgangspunten. De feiten die zich in Nederland voordoen en waar informatiebehoefte over is, kunnen worden gemodelleerd als real-world objecten met eigenschappen die in de loop der tijd kunnen veranderen. Deze feiten worden afgebeeld op een dataset door de waarde van een attribuutvariabele te meten[3].

- De waarde van de *attribuutvariabele* refereert naar een *eigenschap* van een *real-world object* met betrekking tot een bepaalde *periode*:
    - De beschrijving van de variabele met haar waardedomein typeert de eigenschap. Voorbeeld: de eigenschap ’42 jaar’ met als attribuutvariabele ‘leeftijd in jaren’.
    - Het real-world object maakt deel uit van een populatie. De beschrijving van de populatie typeert het object. Voorbeeld: het real-world object ‘Robbert Renssen’ maakt deel uit van een populatie van personen.
    - De periode waar de eigenschap betrekking op heeft, heet de referentieperiode. Voorbeeld: de eigenschap ‘42 jaar’ van het real-world object ‘Robbert Renssen’ heeft als referentieperiode het ‘jaar 2003’.

Bovengenoemde referentiële componenten (object/populatie, attribuutvariabele, referentieperiode) vormen de ***conceptuele metadata*** van een statistisch product. Deze metadata heeft twee rollen:

- Zij is voorschrijvend. Door de metadata te specificeren (ontwerpen), wordt een keuze gemaakt welke feiten worden afgebeeld.
- Zij is beschrijvend. De gespecificeerde metadata wordt gebruikt om de afbeelding (achteraf) te beschrijven en te interpreteren.

De keuze welke feiten wel en welke feiten niet onder de informatievoorziening vallen, wordt in de ontwerpfase van het statistische proces gemaakt. Deze keuze dient enerzijds klantgericht te zijn. Anderzijds dient de keuze rekening te houden met de zowel de CBS-kosten als de maatschappelijke kosten (administratieve lastendruk).

- Statistische producten moeten relevant en samenhangend zijn, een maatschappelijke informatiebehoefte invullen en (aanwijsbare) externe klanten hebben. Soms ligt er een (wettelijke) verordening aan de keuze ten grondslag.
- Statistische producten mogen (iets) minder relevant zijn als het daardoor veel goedkoper wordt om de afbeelding te maken. Het CBS-beleid stelt dat secundaire waarneming de voorkeur verdient boven primaire waarneming (hergebruik data).

---

[3] Voor het maken van de afbeelding zijn productieprocessen ingericht, zie deel C.

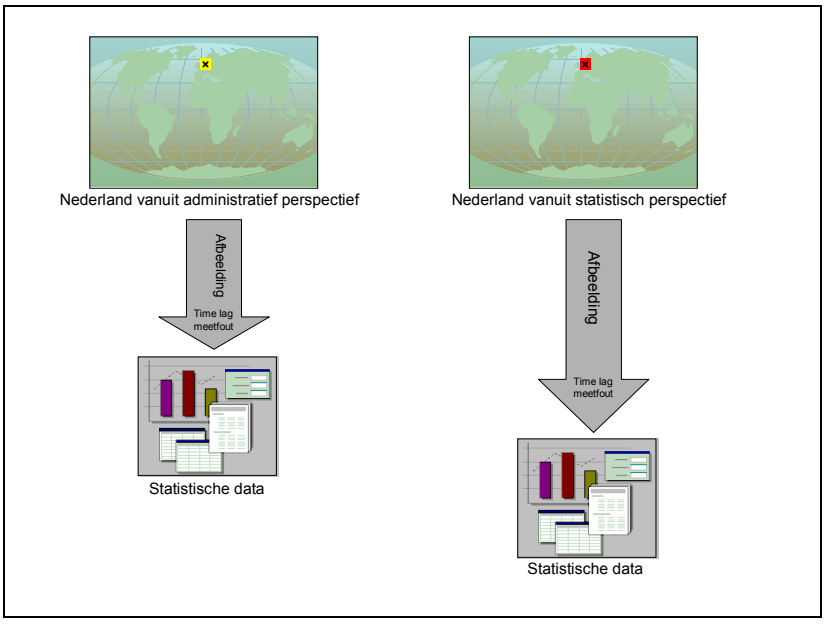
Zo kan er een informatiebehoefte naar werkloosheidcijfers zijn. Om hier een statistisch product van te maken dient de doelpopulatie te worden afgebakend, dient de eigenschap ‘werkloosheid’ te worden gespecificeerd en dient de referentieperiode te worden aangegeven. Ter illustratie wordt de volgende (vereenvoudigde) specificatie over de werkloosheidscijfers gegeven.

- Populatieafbakening: Nederlands ingezetenen conform de Gemeentelijke Basis Administratie (GBA) van de leeftijd 15-65 jaar.
- Specificatie Werkloosheid: Minder dan 12 uur per week werkend en actief op zoek naar (meer) werk.
- Referentieperioden: Maand, kwartaal en jaar.

Merk op dat, gegeven de informatiebehoefte over werkloosheidscijfers, er bij de specificatie van de conceptuele componenten (object/populatie, attribuutvariabele, referentieperiode) ontwerpkeuzes zijn gemaakt. Idealiter wordt daarbij zoveel mogelijk aangesloten op statistische concepten. Dit zijn concepten die vanuit de statistische informatiebehoefte *relevant* en *gecoördineerd* zijn. Daarentegen wordt vanuit kostenoverwegingen zoveel mogelijk aangesloten op de administratieve concepten, i.e. de concepten die gehanteerd zijn in de administraties van de aanleverende partijen van data (veelal registerhouders).

In figuur B-2 is het verschil in waarnemings- en verwerkingstijd schematisch weergegeven aan de hand van de lengte van de pijl van de afbeelding.

Figuur B-2: Administratieve versus statistische invalshoek



Bij de specificatie van de werkloosheid hierboven is gekozen voor het statistische concept, en bijvoorbeeld niet voor het administratieve concept zoals dat is gehanteerd bij het Centrum voor Werk en Inkomen (CWI).

3.1.2 Kwaliteitsmetadata

Naast deze referentiële componenten van een statistisch product zijn er ook kwaliteitscomponenten. De afbeelding kent vaak een zekere time lag en kan bovendien fouten bevatten. Kwaliteitscomponenten zijn zeker zo belangrijk omdat zij samen met de referentiële componenten de inferentiële betekenis van een statistische dataset vastleggen.

De ***kwaliteitsmetadata*** zegt iets over de kwaliteit van de gemaakte afbeelding. Ook deze metadata heeft twee rollen.

- Zij is voorschrijvend. Door de metadata te specificeren (ontwerpen), worden kwaliteitseisen gesteld ten aanzien van time lag (actualiteit) en fout (betrouwbaarheid en/of cijferconsistentie) van de afbeelding.
- Zij is beschrijvend. De gespecificeerde metadata wordt gebruikt om de kwaliteit van de afbeelding (achteraf) te kunnen interpreteren.

Als de afbeeldingen van Nederland worden opgevat als een verzameling foto's dan dienen deze foto's een in elkaar passend (*gecoördineerd, compleet en consistent)* en scherp (*betrouwbaar*) beeld te geven van *relevante* feiten in Nederland[4]. Hierbij is het kwaliteitsaspect *samenhang* in drie deelaspecten gesplitst, namelijk *gecoördineerd*, *compleet* en *consistent*.

De kwaliteitsmetadata kunnen vanuit verschillende gezichtspunten bekeken worden. In Daas en Arends-Tòth (2008) worden deze gezichtspunten hyperdimensies genoemd. De relevantie, compleetheid en coördinatie van de CBS-producten betreffen kwaliteits-aspecten die betrekking hebben op de keuze van de invalshoek van de afbeelding. De cijferconsistentie en betrouwbaarheid zijn kwaliteitsaspecten die betrekking hebben op de data van de afbeelding.

Merk op dat deze deelparagraaf zich heeft beperkt tot kwaliteitsmetadata met betrekking tot de data. In de vorige deelparagaaf (paragraaf 3.1.1) kwam de keuze van de invalshoek aan de orde.

3.2 Logische structuur van een afbeelding

De logische structuur van een afbeelding volgt de wijze waarop de eigenschappen van de real-world objecten geordend dan wel gestructureerd kunnen worden. Voorafgaand aan de logische structuur zullen we daarom eerst structuur aanbrengen in de 'real-world'.

3.2.1 Populaties en klassen

Een populatie is een verzameling van real-world objecten en specificeert daarmee een afbakening. Een ***populatietype*** is een typering van een populatie aan de hand van één of

[4] Vaak worden ook *actualiteit* en *toegankelijkheid* als kwaliteitsaspecten genoemd. Deze aspecten komen in paragraaf 4 aan de orde.

meer eigenschappen. Een voorbeeld van een populatie is {Jantje, Pietje, Klaasje, Mientje, Aaltje}. De Nederlandse bevolking van het jaar 2001 is een voorbeeld van een typering van een populatie. Ieder persoon die in 2001 Nederlander was, behoort tot dit populatietype.

De eigenschappen die een populatie typeren, i.e. de eigenschappen die een object moet bezitten om tot die populatie te behoren noemen we de definiërende eigenschappen van een populatietype.

Naast het begrip *populatietype* kennen we ook het begrip *objecttype*. Ook een ***objecttype*** typeert een populatie, maar niet aan de hand van een tijd- of ruimtecomponent. Voorbeelden van objecttypen zijn 'persoon', 'huishouden', 'rit', 'bedrijfseenheid', etc.

Gegeven een populatie van een bepaald type is het mogelijk om aan de hand van één of meer niet-definiërende eigenschappen deelpopulaties te vormen. Iedere deelpopulatie is op te vatten als een verdere afbakening. Uiteraard zijn de eigenschappen waarmee een deelpopulatie is gevormd definiërend ten aanzien van de deelpopulatie. De populatie {Jantje, Pietje, Klaasje} is een deelpopulatie van {Jantje, Pietje, Klaasje, Mientje, Aaltje}. Het man-zijn is een definiërende eigenschap voor deze deelpopulatie.

- Indien de uit een populatie gevormde deelpopulaties niet-overlappend zijn en samen de populatie vormen dan spreken we van klassen i.p.v. deelpopulaties.
- De populatie waarvan een deelpopulatie een afgrenzing is noemen we de referentiepopulatie m.b.t. die deelpopulatie. Zodra we dus over een deelpopulatie (of klasse) spreken dan is er dus altijd sprake van een referentiepopulatie.

De verzameling van klassen, waarbij iedere klasse is opgevat als real-world object, kan weer worden opgevat als een populatie. Uiteraard bezitten ook dergelijke geclusterde real-world objecten eigenschappen die kunnen worden gemeten. Als voorbeeld onderscheiden we in de populatie {Jantje, Pietje, Klaasje, Mientje, Aaltje} twee klassen: {Jantje, Pietje, Klaasje} en {Mientje, Aaltje}. Deze twee klassen vormen samen een populatie van klassen waarop we de attribuutvariabele 'omvang' kunnen definiëren.

3.2.2 Classificerende en kwantificerende attribuutvariabelen

Een beetje abstract geredeneerd kan een attribuutvariabele worden opgevat als een functie die eigenschappen van een populatie van real-world objecten afbeeldt op een getallenset (waardedomein). Bij die afbeelding vindt feitelijk een meting plaats waarbij meerdere meetschalen kunnen worden onderscheiden:

- De nominale schaal: de afgebeelde getallen zijn kengetallen en uitsluitend bedoeld om de (real-world) objecten te classificeren.
- De ordinale schaal: de afgebeelde getallen zijn ranggetallen waarmee de (real-world) objecten kunnen worden geclassificeerd en geordend.

- De ratioschaal: de afgebeelde getallen zijn complete maatgetallen waarmee de (real-world) objecten kunnen worden geclassificeerd, geordend en naar ratio onderling vergeleken.

Uit bovenstaande meetschalen blijkt dat met iedere attribuutvariabele real-world objecten in klassen kunnen worden ingedeeld, maar dat niet met iedere attribuutvariabele kan worden gerekend. Indien een attribuutvariabele wordt gebruikt om een populatie in klassen te delen dan spreken we van een ***classificerende variabele***. Samen met de definitie van de populatie, definieert een dergelijke variabele de klassen.

Deze klassen kunnen ook weer eigenschappen hebben. Indien een eigenschap van een klasse is afgebeeld door een totaal of gemiddelde te nemen van een attribuutvariabele welke de eigenschappen van elk van de afzonderlijke objecten afbeeldt, dan spreken we van een ***kwantificerende variabele.*** Merk op dat het niet louter om totalen of gemiddelden hoeft te gaan, maar dat ook complexere eigenschappen van een klasse kunnen worden afgebeeld.

Tabel 1: Voorbeeld van een statistische tabel

Huishoudens; gemiddeld inkomen

Kenmerken	Totaal particuliere huishoudens	Perioden	2002*

Onderwerpen	Personen per huishouden	Besteedbaar inkomen per huishouden
Samenstelling huishouden	*absoluut*	*1000 euro*
Eenpersoonshuishouden	1,0	18,4
Meerpersoonshuishouden	2,9	34,2

© Centraal Bureau voor de Statistiek, Voorburg/Heerlen 2005-03-21

Ter illustratie beschouwen we een populatie van particuliere huishoudens in Nederland, afgebakend naar het jaar 2002. Op deze huishoudens zijn de volgende eigenschappen gemeten:

- Het aantal personen in het huishouden (één, twee, drie,…) in het jaar 2002.
- De samenstelling van het huishouden (één of meerpersoons) in het jaar 2002, welke kan worden afgeleid uit het aantal personen per huishouden in dat jaar.
- Het besteedbare inkomen van het huishouden (in 1000 Euro's) in het jaar 2002.

Merk op dat het aantal personen in een huishouden en het besteedbare inkomen in de ratioschaal zijn gemeten, terwijl de samenstelling van het huishouden een meting betreft in de ordinale schaal. Op basis van de samenstelling van het huishouden kan de populatie

van huishoudens gesplitst worden in twee klassen: de klasse van éénpersoonshuishoudens en de klasse van meerpersoonshuishoudens.

Tabel 1 geeft een statistische tabel waarin op beide klassen twee eigenschappen zijn gemeten, namelijk het gemiddelde aantal personen van de huishoudens waaruit de klasse bestaat en het gemiddelde besteedbaar inkomen van de huishoudens waaruit de klasse bestaat. In deze tabel is de attribuutvariabele 'samenstelling huishouden' gebruikt als classificerende variabele en zijn de variabele 'besteedbaar inkomen' en 'aantal personen' gebruikt als kwantificerende variabelen.

3.2.3 Relaties tussen object- c.q. populatietypen

Zoals gezegd karakteriseert een populatietype een populatie aan de hand van een set van definiërende eigenschappen, in die zin dat alle real-world objecten die deze eigenschappen bezitten tot de populatie behoren.

Indien nu aan deze set van definiërende eigenschappen een eigenschap wordt toegevoegd dan ontstaat als het ware een nieuwe karakterisering. Ten opzichte van het oorspronkelijke populatietype spreken we van een populatiesubtype. Een voorbeeld van een populatietype is de Nederlandse bevolking uit 2002 met als subtype daaruit de Nederlandse mannelijke bevolking uit 2002.

Op dezelfde wijze kan uit de set van definiërende eigenschappen een eigenschap worden weggelaten. Ten opzichte van het oorspronkelijke populatietype spreken we dan van een populatiesupertype. Ten opzichte van de Nederlandse bevolking 2002 is de wereldbevolking uit 2002 een voorbeeld van een populatiesupertype.

Een andere vorm waarop real-world objecten aan elkaar gerelateerd zijn is dat zij opgebouwd kunnen zijn uit componenten en de componenten zelf ook weer als real-world objecten kunnen worden opgevat. Een huishouden bestaat uit personen en beide kunnen als een real-world object worden opgevat met een relatie daartussen. Een persoon is in dit voorbeeld een componentobject van het object huishouden.

Een derde vorm van een relatie tussen real-world objecten ontstaat als een object een relatie is tussen twee andere objecten. Een voorbeeld van een relatie-object is het dienstverband tussen persoon A en bedrijf B.

3.2.4 Logische structuur van een afbeelding

Tabel 2 geeft een voorbeeld van een (fysieke) statistische dataset. De dataset is gemakshalve gepresenteerd als een bestand met op de rijen (unieke) identificatienummers (deze verwijzen naar real-world objecten in Nederland en worden aangegeven met sleutelvariabelen) en op de kolommen de attribuutvariabelen (de waarden van de attribuutvariabelen verwijzen naar de eigenschappen van de real-world objecten).

In tabel 2 kunnen twee objecttypen worden onderscheiden, namelijk huishouden en persoon. Een huishoudens-id identificeert een huishouden en de combinatie van een

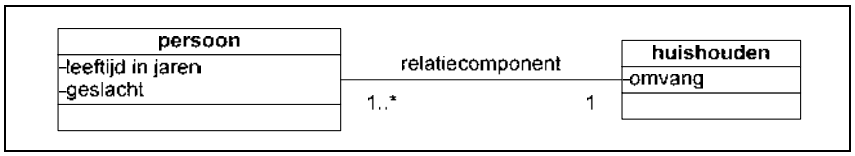
huishoudens-id met een volgnummer identificeert een persoon. Er zijn twee attribuutvariabelen waarvan de waarden naar eigenschappen van personen refereren. Dit zijn ‘leeftijd’ en ‘geslacht’. Er is één attribuutvariabele waarvan de waarden naar eigenschappen van huishoudens refereren. Dit is ‘omvang huishouden’.

Tabel 2: Voorbeeld van een statistische dataset (micro)

Sleutelvariabelen		**Attribuutvariabelen**		
Huishoudens-id	Volgnummer	Leeftijd in jaren	Geslacht	Omvang huishouden
1	1	40	V	3
1	2	43	M	3
1	3	20	V	3
2	1	25	V	2
2	2	26	M	2
3	1	70	M	1
4	1	50	M	4
…				…

De logische structuur die uit de fysieke presentatie van tabel 2 gehaald kan worden is weergeven is figuur B-3. De logische structuur van figuur B-3 kent twee dimensies, een persoonsdimensie en een huishoudensdimensie. Voor elk van de dimensies is er een relatiecomponent die de personen en huishoudens aan elkaar relateert.

Figuur B-3: Logische structuur van een statistische dataset



Tabel 2 is een voorbeeld van een statistische dataset waarbij de objecten op ‘micro niveau’ zijn, in die zin dat de real-world objecten waarover de tabel gaat via sleutel-id’s onderling van elkaar onderscheiden zijn. Een voorbeeld van een statistische dataset op geaggregeerd niveau is tabel 1. Gemakshalve is deze tabel in dezelfde vorm gegoten als de micro dataset, zie tabel 3.

Tabel 3: Voorbeeld van een statistische dataset (geaggregeerd)

Classificerende variabelen	**Kwantificerende variabelen**	
Samenstelling huishouden	Gemiddeld aantal personen per huishouden	Gemiddeld Besteedbaar inkomen (in 1000 Euro’s) per huishouden
Eenpersoons	1	16,4
Meerpersoons	2,9	34,2

In tabel 3 kan één objecttype worden onderscheiden. Dit objecttype betreft klassen van huishoudens, waarbij de klassen zijn gedefinieerd aan de hand van de classificerende variabele 'samenstelling huishouden'. De klassen kunnen worden opgevat als (aggregaat) object en de huishoudens als componentobject.

Er zijn twee kwantificerende variabelen die een eigenschap van deze klassen afbeelden, namelijk 'gemiddeld aantal personen per huishouden' en 'gemiddeld besteedbaar inkomen (in 1000 Euro's) per huishouden'.

Laat X_i het totaal aantal personen in klasse i zijn, Y_i het totaal aantal huishoudens in klasse i, en Z_i het totaal besteedbare inkomen in klasse i (i=1,2), zie tabel 4. Dan is het gemiddelde aantal personen per huishouden gedefinieerd als Y_i / X_i en is het gemiddeld besteedbare inkomen als Z_i / X_i .

Tabel 4: Vervolgvoorbeeld van een statistische dataset (geaggregeerd)

Classificerende variabelen	**Kwantificerende variabelen**		
Samenstelling huishouden	Totaal aantal personen	Totaal aantal huishoudens	Totaal besteedbaar inkomen (in 1000 Euro's)
Eenpersoons	X_1	Y_1	Z_1
Meerpersoons	X_2	Y_2	Z_2

Een vergelijking tussen tabel 2 enerzijds en tabel 3 of 4 anderzijds laat zien dat de sleutelvariabelen die de micro-objecten aangeven, vervangen zijn door classificerende variabelen die de geaggregeerde objecten aangeven. Verder wordt in tabel 3 en 4 niet langer gesproken over attribuutvariabelen, maar over kwantificerende variabelen.

3.3 Logisch versus fysiek datamodel van een afbeelding

In deze cursus wordt onder een ***logisch datamodel*** verstaan het model dat de logische structuur van een dataset beschrijft, zie bijvoorbeeld figuur B-3. In dit model worden de objecttypen, de bij de objecttypen behorende variabelen en de relaties tussen de objecttypen benoemd. Indien het model uit één objecttype bestaat dan spreken we van een enkelvoudig model. Bij twee of meer objecttypen spreken we van een complex model.

Vaak kan een statistische dataset met een bepaalde logische structuur op meerdere manieren fysiek vastgelegd worden, eventueel verspreid over meerdere bestanden. Voortbordurend op de logische structuur van tabel 2 bespreken we drie fysieke structuren:

- De eerste fysieke structuur bestaat uit één rechthoekige vorm waarbij de rijen worden gevormd door de persoonsnummers en de kolommen door de variabelen 'huishoudensnummer', 'leeftijd in jaren', 'geslacht' en 'omvang huishouden'. Dit is feitelijk de fysieke structuur zoals gegeven in tabel 2. Nadeel van deze structuur is dat 'omvang huishouden' voor iedere persoon in een huishouden herhaald wordt.

- De tweede fysieke structuur bestaat eveneens uit een rechthoekig bestand, maar nu zijn de rijen gevormd door de huishoudensnummers en de kolommen door de variabelen ‘omvang huishouden’, ‘persoonsnummer van persoon X’, ‘leeftijd in jaren van persoon X’, en ‘geslacht van persoon X’, waarbij X loopt van 1 t/m (zeg) 20. Merk op dat bij deze structuur ervan wordt uitgegaan dat alle huishouden uit minder dan 20 personen bestaat. Nadeel van de structuur is dat voor de huishoudens die uit minder 20 personen bestaan – dit zijn de meeste – er vele lege velden zullen zijn en dat huishoudens met meer dan 20 personen zijn afgeknot.
- De derde fysieke structuur bestaat uit drie rechthoekige bestanden; een rechthoekig bestand met op de rijen de persoonsnummers en op de kolommen de variabelen ‘leeftijd in jaren’ en ‘geslacht’, een rechthoekig bestand met op de rijen de huishoudensnummers en op de kolommen de variabele ‘omvang huishouden’, en een rechthoekig bestand waarin de persoonsnummers en de huishoudnummers aan elkaar gerelateerd zijn.

Alledrie fysieke structuren bevatten inhoudelijk dezelfde statistische informatie en zijn in elkaar over te zetten. Het is veelal een kwestie van smaak, van opslagcapaciteit of van toolkeuze welke structuur het handigst in gebruik is.

In deze cursus wordt onder een ***fysiek datamodel*** verstaan het model dat de fysieke structuur van een dataset beschrijft. In dit model worden de betrokken bestanden beschreven met betrekking tot de positie en syntax van de in de logische structuur voorkomende variabelen, als ook het formaat van de bestanden inclusief de gehanteerde tekenset.

Gegeven de logische structuur van een dataset, kan het ***technisch (niet-inhoudelijk) transformeren*** nu omschreven worden als het omzetten van een dataset van de éne fysieke structuur naar een andere fysieke structuur.

3.4 Geldigheidsperiode van een afbeelding

Er zijn twee redenen waarom een afgebeeld feit in een dataset een beperkte geldigheid heeft.

- Ten eerste omdat het feit zelf, i.e. het real-world object met eigenschap, in de loop der tijd kan veranderen. In dit geval spreken we van een ***reële mutatie***.
- Ten tweede omdat we erachter komen dat het feit verkeerd is afgebeeld en we deze fout willen herstellen. Bij foutherstel spreken we van een ***niet-reële mutatie***.

3.4.1 Reële mutaties

Stel dat in het eerste huishouden van tabel 1 een verandering per 1-1-2007 optreedt doordat de derde persoon in dit huishouden een zelfstandig huishouden begint. En stel verder dat deze mutatie is waargenomen in de vorm van een mutatiebericht dat ons (het CBS) bereikt heeft op 3-1-2007 (er is enige vertraging; time lag).

Tabel 5: Voorbeeld van een statistische dataset (geldig vanaf 1-1-2007)

Sleutelvariabelen		Attribuutvariabelen		
Huishoudens-id	Volgnummer	Leeftijd in jaren	Geslacht	Omvang huishouden
1	1	40	V	2
1	2	43	M	2
2	1	25	V	2
2	2	26	M	2
3	1	70	M	1
4	1	50	M	4
5	1	20	V	1
…				…

Als we de mutatie in de dataset willen verwerken dan heeft dit drie gevolgen. Ten eerste zal er een nieuw huishouden opgevoerd moeten worden. Dit is een verandering van de waarde van een sleutelvariabele, immers er komt een waarde bij. Ten tweede zal de omvang van het eerste huishouden afnemen van 3 personen naar 2 en zal de omvang van het nieuwe huishouden op 1 persoon vastgesteld worden. Dit is een verandering respectievelijk opvoering van de waarde van een attribuutvariabele. Ten derde zal de relatie tussen het eerste huishouden en de derde persoon in dit huishouden ophouden te bestaan en zal er een relatie tussen het nieuwe huishouden en deze persoon ontstaan. In tabel 5 is de situatie na de verandering afgebeeld.

De situatie van tabel 2 was geldig (vlak) voor 1-1-2007 en de situatie van tabel 5 (vlak) na 1-1-2007. Door tabel 2 en 5 met elkaar te vergelijken kunnen de verschillen tussen twee standen ten aanzien van de personen en de huishoudenspopulatie worden waargenomen. Wat overigens niet in tabel 5 kan worden geconstateerd, is dat er een relatie is tussen het eerste en vijfde huishouden. Geen van beide tabellen beschrijven dan ook ouder-kindrelaties als objecttype.

3.4.2 Foutherstel

Stel dat een tweede mutatiebericht het CBS bereikt (op 4-1-2007) met de mededeling dat het eerste mutatiebericht ten onrechte was. De derde persoon in het eerste huishouden is niet een eigen huishouden begonnen. De stand van zaken volgens het laatste bericht is wederom volgens tabel 2. Conform de mutatieberichten geldt de volgende reeks van constateringen:

- de omvang van huishouden 1 is gelijk aan 3 tot 1 januari 2007 (opgevat als een waar feit tot 4 januari)
- de omvang van huishouden 1 is gelijk aan 2 vanaf 1 januari 2007 (opgevat als een waar feit tot 4 januari).

- de omvang van huishouden 1 is gelijk aan 3 ook vanaf 1 januari 2007 (opgevat als een waar feit vanaf 4 januari).

Afhankelijk van het moment waarop de mutatieberichten zijn verwerkt en waarop de afbeeldingen worden opgevraagd hebben we met één van deze drie afbeeldingen te maken. Indien de stand van 2 januari 2007 wordt aangevraagd zoals geregistreerd op 2 januari 2007 dan wordt tabel 2 verkregen. Wordt dezelfde stand aangevraagd maar nu zoals geregistreerd op 3 januari dan wordt tabel 5 verkregen. Wordt deze stand op 4 januari aangevraagd dan is deze weer conform tabel 2.

3.4.3 Geldigheids- versus referentieperioden

Samenvattend kan worden gesteld dat een afgebeeld feit altijd over een referentieperiode gaat en dat vanwege een reële mutatie of foutherstel de afbeelding voor een zekere periode geldig is. Het is van belang deze verschillende rollen van perioden van elkaar te onderscheiden:

- een referentieperiode waar een reëel feit, i.e. een eigenschap van een real-world object, betrekking op heeft. De referentieperiode verandert zodra zich een reële mutatie voordoet.
- een geldigheidperiode waar de afbeelding van een reëel feit betrekking op heeft. De geldigheidsperiode verandert zodra er een reële mutatie is of een foutherstel.

Op basis van beide rollen (dual time) moet een gebruiker van statistische data in staat zijn om een statistische dataset te verkrijgen dat betrekking heeft op referentieperiode A, maar waarvan er, afhankelijk van de momenten waarop foutherstel plaatsvindt, meerdere (kwaliteits)versies zijn.

Het resultaat is een afbeelding van een populatie voor een referentieperiode A, die vanwege (nog) niet herstelde fouten, iets kan afwijken van de beoogde populatie. De beoogde populatie is vastgelegd in de referentiële betekenis. Kennis over de afwijking is nodig voor de inferentiële betekenis.

3.5 Statistische data, statistische gegevens en statistische informatie

Samenvattend kunnen we stellen dat een statistisch product uit een aantal componenten bestaat, namelijk een dataset met een fysieke structuur, een logisch datamodel behorende bij de fysieke structuur, een inhoudelijke (conceptuele) beschrijving behorende bij het logische model, en een kwaliteitsbeschrijving. De genoemde meta-componenten zijn nodig om de fysieke dataset te kunnen begrijpen. In deze cursus onderscheiden we

- Statistische data: dataset met fysiek datamodel zonder logisch datamodel en zonder beschrijvende conceptuele metadata en kwaliteitsmetadata.
- Statistische gegevens: dataset met fysiek datamodel met een logisch datamodel en met beschrijvende conceptuele metadata en kwaliteitsmetadata.

Er kan gesteld worden dat de conceptuele metadata en de kwaliteitsmetadata de statistische data naar een hoger kennisniveau tillen.

Een volgend kennisniveau wordt verkregen door de statistische gegevens te duiden (analyseren), te presenteren (visualiseren) en te communiceren (uitleggen via toelichtingen en/of artikelen). Daarbij is het mogelijk om een onderscheid te maken tussen verschillende doelgroepen, waarbij iedere doelgroep een 'eigen' duiding met 'eigen' uitleg en presentatievorm krijgt. Deze doelgroep specifieke interpretatie van statistische gegevens wordt in deze cursus statistische informatie genoemd.

- Statistische informatie: statistische gegevens met een duiding, een presentatievorm en een uitleg.

We sluiten het hoofdstuk af door te stellen dat het CBS drie typen van eindproducten heeft waarvoor dus externe klanten zijn: statistische informatie (een voorbeeld is tabel 1) en statistische gegevens met een onderscheid tussen microgegevens (een voorbeeld is tabel 2) en geaggregeerde gegevens (een voorbeeld is tabel 3).

4. Praktijkvoorbeeld: de statistiek van de binnenvaart

4.1 Doel en kort overzicht van de binnenvaart

De statistiek van de binnenvaart geeft informatie over het binnenlandse en internationale goederenvervoer en verkeersbewegingen over de binnenwateren. De belangrijkste outputverplichtingen van de statistiek van de binnenvaart worden benoemd in paragraaf 4.2.

De belangrijkste inputbronnen voor de statistiek van de binnenvaart vormen de IVS-bestanden. IVS staat voor Informatie- en Volgsysteem Scheepsvaart en is een door Rijkswaterstaat opgezet computernetwerk van belangrijke telpunten op de hoofdvaarwegen in Nederland (sluizen, verkeersposten). Op deze telpunten worden gegevens ingewonnen over het passerende schip, de vervoerde lading en de reis.

Het systeem dient voornamelijk als hulpmiddel voor de vaarwegbeheerders om het scheepvaartverkeer in goede banen te leiden en zo een vlot en veilig transport over water te garanderen. De informatie van de verschillende IVS-punten wordt dagelijks beschikbaar gesteld aan het CBS.

Voor de reizen die niet langs IVS telpunten voeren verkrijgt het CBS aanvullende bronnen[5]. Zo dienen schippers op dergelijke vaarroutes een formulier in te vullen (een

[5] De schippers zijn conform de Wet Goederenvervoer Binnenvaart (WGB) verplicht opgave te doen aan het CBS van de door hun gemaakt reizen in een maand. Om de enquêtedruk te verminderen en dus de schippers te ontlasten maakt het CBS zoveel mogelijk gebruik van de IVS-gegevens. Voor vaarroutes waar (nog) geen IVS-punten zijn, blijft de traditionele dataverzameling van kracht.

zogenaamde maandstaat) en deze op te sturen naar het CBS. Dergelijke aanvullende bronnen vormen echter een klein aandeel van de totale gegevensinzameling.

Kortom, het CBS verkrijgt via het IVS-systeem en de aanvullende bestanden een min of meer integrale telling met betrekking tot gemaakte reizen door de binnenschepen. Hier en daar wordt er een reis gemist of zijn de gegevens over een reis niet compleet of foutief ingevuld.

Het **doel** van de statistiek van de binnenvaart is om op basis van deze verkregen inputs over reizen op efficiënte wijze aan de geformuleerde outputverplichtingen te kunnen voldoen. Simpel gezegd komt deze wijze neer op 1) het controleren van de verkregen reisinformatie op gaten, fouten en overlap, 2) het herstellen van de geconstateerde gebreken en 3) het deductief afleiden van de outputvariabelen. Het resultaat hiervan is een gaafgemaakt 'integraal' microbestand dat de basis vormt voor de gewenste output. Om bijvoorbeeld de gedefinieerde outputtabellen te 'schatten' hoeven er alleen nog maar een selectie- en een (ongewogen) aggregatieslag op het gaafgemaakte bestand plaats te vinden.

4.2 Statistische gegevens bij de binnenvaart

4.2.1 Opsomming en logische structuur

De binnenvaart levert eindproducten aan verschillende 'leveranciers'. Aan Eurostat worden de volgende producten geleverd (verordening (EG), nr. 1365/2006):

- A1: Goederenvervoer naar soort goederen (jaargegevens)
- B1: Vervoer naar scheepsvlag en soort schip (jaargegevens)
- B2: Scheepsverkeer (jaargegevens) t.b.v. verkeersstatistiek
- C1: Containervervoer naar soort goederen (jaargegevens)
- D1: Vervoer naar scheepsvlag (kwartaalgegevens)
- D2: Containervervoer naar scheepsvlag (kwartaalgegevens)
- E1: Goederenvervoer (jaargegevens)

Daarnaast worden er eindproducten aan Havenbedijf Rotterdam, Statline, CCR, DVS en NIS 'opgeleverd'. Dit zijn de volgende tabellen

- Pubbestand Binnenlands
- Pubbestand Internationaal
- Outputproduct (CCR)
- Outputproducten i.v.m. grensovergang België (NIS: Kreekrak, Wessem, Zelzate)

- ‘Terugmelding DVS’ met een ‘IVS-layout’.
- Statline tabellen: scheepvaartbeweging, goederenstromen, grensovergang en goederenoverslag.

Ter illustratie worden in appendix A de logische datamodellen van de leveringen aan Eurostat gegeven. Op dezelfde wijze kunnen de logische datamodellen van de andere eindproducten worden opgesteld.

Zoals gezegd kent de statistiek van de binnenvaart één omvangrijke hoofdstroom die haar voorziet van de benodigde input, namelijk ‘IVS-nieuw’. Deze stroom betreft momenteel 11 bestanden die dagelijks via EDV (Elektronische DataVerkeer) aan het CBS worden geleverd door het bedrijf SPITS. Naast de hoofdstroom kent de binnenvaart vier aanvullende stromen, de maandstaten, de sluisbestanden van Zaandam, die van Vlissingen en aanvullende brug- en sluisformulieren (diskettes) cq bestanden. Ter illustratie is in appendix B het logische datamodel van een IVS-levering gegeven.

De binnenvaart onderkent één belangrijk tussenproduct, namelijk het gaafgemaakte ‘integrale’ microbestand. Het logische datamodel van dit tussenproduct is gegeven in appendix C. Uit dit model blijkt dat het ‘integrale’ microbestand over reizen, zendingen, schepen, containers en telpunten gaat.

Incidenteel komt het voor dat er ‘grove’ fouten in de leveringen zitten die ‘te’ laat ontdekt worden. Om deze reden is besloten om voor de hierboven genoemde producten twee releasemomenten te onderscheiden, namelijk een voorlopige en een definitieve release. Mochten er geen incidenten zijn geweest, dan gaat de voorlopige release van een bepaalde van te voren vastgestelde periode ‘automatisch’ over in de definitieve release.

4.2.2 Inhoudelijke betekenis

De inputbestanden bij de binnenvaart hebben een vorm zoals beschreven in het logische datamodel, zie appendix B. Aan de hand van dit datamodel wordt duidelijk dat de IVS-bestanden over passages, zendingen en containers gaan.

- Een ***passage*** is een geregistreerde oversteek van een vervoerseenheid bij een telpunt;
    - Een ***vervoerseenheid*** is een combinatie van één of meer voertuigen die als eenheid een vervoersbeweging maken.
    - een ***voertuig*** is een transportmiddel
    - een ***telpunt*** is een locatie waar verkeers- en vervoersgegevens geregistreerd worden.
- Een ***zending*** is een bij elkaar horende hoeveelheid van één goed dat in één vervoersbeweging met één voertuig wordt verplaatst van plaats A (lading) naar plaats B (lossing).
- Een ***container*** is een multimodaal verpakkingsmiddel voor transport.

Nu is het zo dat in de inputbestanden van de statistiek van de binnenvaart niet alle passages, alle zendingen, alle containers en alle reizen die ooit in deze wereld hebben plaatsgevonden of hebben bestaan zijn opgenomen. De definities zijn expres generiek gehouden zodat zij ook voor de andere verkeers- en vervoersstatistieken van toepassing zijn. Voor al deze statistieken – en dus ook voor de binnenvaart – moet er dan nog wel een specifieke verdere afbakening komen. Voor de binnenvaart gelden de volgende afbakeningen.

- Een ***passage bij de binnenvaart*** is een geregistreerde oversteek van een vaareenheid bij een binnenvaarttelpunt;
    - Een ***binnenvaarttelpunt*** is een brug, sluis of verkeerspost op Nederlandse binnenwateren waar verkeers- en vervoersgegevens worden verzameld.
    - Een ***vaareenheid*** is een combinatie van één of meer schepen die als samenvarende eenheid een vervoersbeweging maken.
    - Een ***schip*** is een voertuig dat ontwikkeld en/of samengesteld is voor transport over water, dan wel om assistentie te verlenen bij transport over water.
- Een ***zending bij de binnenvaart*** is een zending verplaatst door een schip (onder meer) over de Nederlandse binnenwateren, m.u.v. zeeschepen boven de 5000 ton laadvermogen die niet via de Rijn richting Duitsland gaan.
- Een ***container bij de binnenvaart*** is een zeecontainer van 20-, 30, 40 voet en groter.

Merk op dat de afbakening niet in de tijd is. In principe behoren alle binnenvaartpassages, alle binnenvaartzendingen, alle binnenvaartcontainers en alle binnenvaartreizen tot de doelpopulatie. Echter, vanwege onze menselijke beperkingen in de tijd zijn we niet in staat dit alles waar te nemen en zullen de grondstoffen in de inputbase van een bepaald moment ‘groeien’ in de tijd.

Het gaafgemaakte ‘integrale’ microbestand bij de binnenvaart heeft een vorm zoals beschreven in het logische datamodel, zie appendix C. Aan de hand van dit datamodel wordt duidelijk dat dit microbestand over reizen, schepen, zendingen, containers en telpunten gaat. Afgezien van *reizen* zijn de deze objecttypen plus de specifieke afbakeningen voor de binnenvaart al gedefinieerd:

- Een ***reis*** is een reeks van één of meer vervoersbewegingen van een vaareenheid gekenmerkt door een begin- en eindpunt en een toestand van belading van de schip/schepen waaruit de vaareenheid bestaat.
- Een ***reis bij de binnenvaart*** is een reis van een schip over (onder meer) de Nederlandse binnenwateren, m.u.v. zeeschepen boven de 5000 ton laadvermogen die over het kanaal Gent-Terneuzen naar België gaan.

Ter illustratie staan de definities van de bij alle objecttypen behorende variabelen in appendix D.

De meeste eindproducten bij de binnenvaart zijn tabellen, welke opgebouwd zijn aan de hand van een objecttype en een afbakening daarop, één of meer kwantificerende (intel)variabelen waarvan de waarden in de cellen van de tabel verschijnen en één of meer classificerende variabelen die de randen van de tabel definiëren. Ter illustratie wordt de conceptuele metadata van A1 besproken. Het objecttype van deze tabel is de zending.

Dit objecttype is (in de ruimte) afgebakend door alleen zendingen te beschouwen door schepen over de Nederlands binnenwateren, m.u.v. zeeschepen boven de 5000 ton laadvermogen. Dit objecttype is (in de tijd) verder afgebakend door alleen zendingen te beschouwingen die betrekking hebben op het jaarverslag van de tabel. Van deze doelpopulatie van zendingen worden twee aspecten geschat, namelijk 'vervoerd gewicht' en 'lading tonkilometers'. Dit zijn de kwantificerende intelvariabelen. Wat deze precies inhouden staat bij de inhoudelijke definitie:

- Vervoerd gewicht: opgegeven gewicht van een zending (in tonnen; 1000 kg).
- Lading tonkilometer (Ned): product van het vervoerd gewicht van een zending (in tonnen) en de afgelegde afstand van deze zending (in Nederland).

Deze aspecten worden niet alleen voor de doelpopulatie als geheel gegeven, maar ook per klasse binnen deze populatie, waarbij de klassen gedefinieerd zijn aan de hand van vier classificerende variabelen

- Plaats lading: plaats waar het goed van de zending wordt geladen conform de UNLO-codering, geaggregeerd op regionaal niveau
- Plaats lossing: plaats waar het goed van de zending wordt gelost conform de UNLO-codering, geaggregeerd op regionaal niveau
- Goederensoort: soort goed (de HS-codering); afgeleid uit product
- Verpakkingssoort: toestand van een zending, geclassificeerd naar wel of niet in containers; afgeleid uit verschijningsvorm.

4.3 Statistische data versus statistische gegevens.

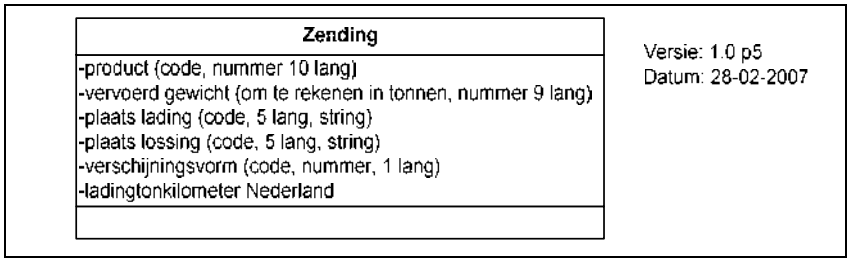
In deze paragaaf willen we ter afsluiting heel kort het verschil tussen statistische data, statistische gegevens en statistische informatie illustreren. Daartoe concentreren we ons op de Eurostattabel A1 (zie appendix A) en het 'integrale' gaafgemaakte microbestand (zie appendix C).

Merk daarbij op dat tabel A1 is verkregen door op het gaafgemaakt 'integrale' microbestand de juiste selectie te maken en op deze selectie een aantal deterministische afleidingen en aggregaties uit te voeren. Het logische datamodel van A1 is gegeven in tabel 7 en het logische datamodel van de selectie in tabel 8.

Tabel 7: logisch datamodel van A1

Objecttype	*Afbakening*	*Classificerende variabele*	*Kwantificerende variabele*
zending	Nederlandse binnenwateren	- plaats lading - plaats lossing - goederensoort - verpakkingssoort	vervoerd gewicht lading tonkilometers (Ned)

Tabel 8: logische datamodel van een selectie op integraal microbestand



De geaggregeerde (fysieke) data bij het logische datamodel van A1 en de (fysieke) microdata bij het logische datamodel van de selectie zijn statistische data. Bij het technische datamodel van deze data zijn vaak wel de variabelennamen benoemd – deze komen overeen met de variabelennamen in tabel 7 en 8 – maar er zijn geen inhoudelijke beschrijvingen gegeven. De betekenis van de data wint aan kracht indien duidelijk wordt wat onder een *zending* wordt verstaan en wat onder de verschillende *variabelen* wordt verstaan:

- Een ***zending*** is een bij elkaar horende hoeveelheid van één goed dat in één vervoersbeweging met één voertuig wordt verplaatst van plaats A (lading) naar plaats B (lossing).
- ***Vervoerd gewicht***: opgegeven gewicht van een zending (in tonnen; 1000 kg).
- ***Lading tonkilometer*** (Ned): product van het vervoerd gewicht van een zending (in tonnen) en de afgelegde afstand van deze zending (in Nederland).
- ***Plaats lading***: plaats waar het goed van de zending wordt geladen conform de UNLO-codering.
- ***Plaats lossing***: plaats waar het goed van de zending wordt gelost conform de UNLO-codering.
- ***Goederensoort***: soort goed (conform de HS-codering)
- ***Verpakkingssoort***: toestand van een zending, geclassificeerd naar wel of niet in containers.

De betekenis zou verder aan kracht winnen indien duidelijk is wat onder de UNLO- en HS-codering wordt verstaan. De betekenis zou nog verder aan kracht winnen als er informatie was over de betrouwbaarheid.

Deel C: Het Statistische Proces

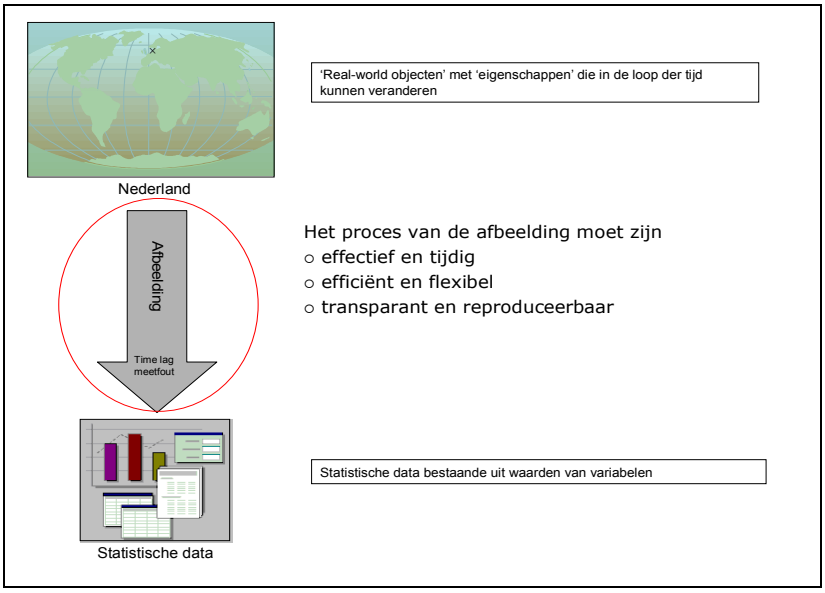
Dit deel gaat in op de afbeelding als statistisch proces, zie figuur C-1. Ten behoeve van de afbeelding, waarin een ontworpen statistisch eindproduct voor externe klanten wordt gerealiseerd, is veelal een procesketen ingericht, waarbij

- meerdere organisatieonderdelen betrokken zijn die verantwoordelijk zijn voor een deel van de keten,
- en waarbinnen ieder deel van de keten (en dus binnen de organisatieonderdelen) één of meer processen zijn ingericht.

De ketengedachte en de daarbij behorende uitwisseling van de statistische tussen- en eindproducten is het onderwerp van paragraaf 5.

Bij het realiseren van de tussen- en eindproducten zijn van oudsher drie fasen te onderkennen, namelijk input (waarnemen), throughput (verwerken en statistische beveiligen) en output (informatieontwikkeling en publicatie). In paragraaf 6 worden deze fasen kort geïntroduceerd en in paragraaf 7 tot en met 10 uitgewerkt.

Figuur C-1: Het proces van de afbeelding



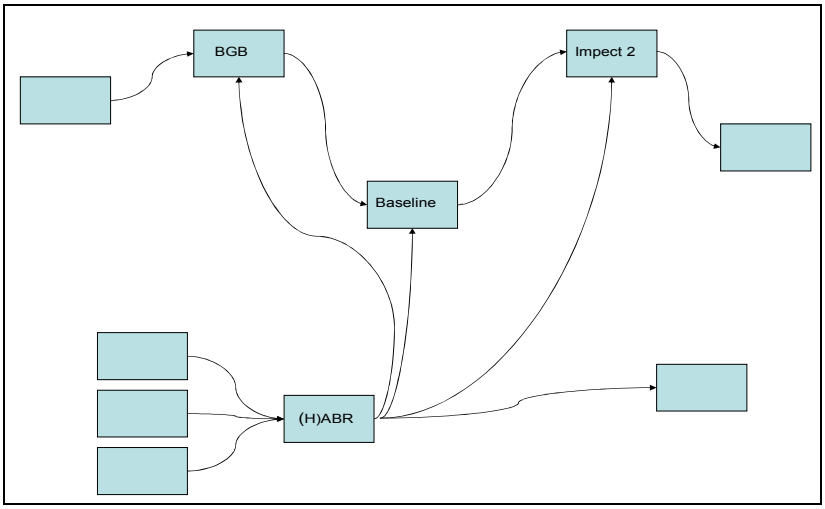
5. Het statistische proces als keten met schakels

5.1 Voorbeeld

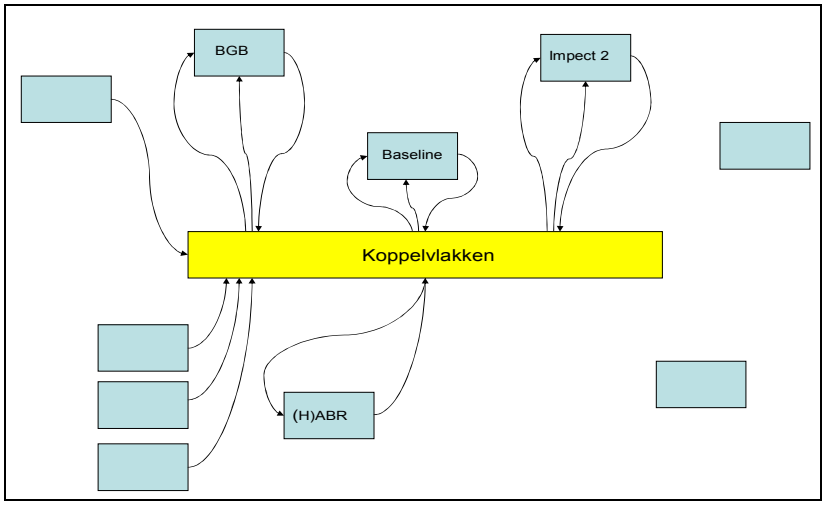
Aansluitend op de huidige praktijk van het CBS kan gesteld worden dat veel processen geautomatiseerd zijn en dat deze processen herkenbaar zijn aan de naamgeving van het automatiseringssysteem. Bij de KS met BTW-omzetgegevens kunnen we bijvoorbeeld spreken van processen rondom BGB, processen rondom Baseline en processen rondom

Impect II[6]. De in- en outputs van deze systemen vormen ten aanzien van de omzetgegevens een waardeketen, zie Renssen en Kruiskamp (2008).

Figuur C-2: netwerk tussen verwerkingssystemen (zonder koppelvlak)



Figuur C-3: netwerk tussen verwerkingssystemen (met koppelvlak)



Om deze omzetgegevens te kunnen verwerken zijn vaak nog andere gegevens nodig. Bij bijvoorbeeld de hierboven genoemde verwerkingssystemen zijn er gegevens nodig uit de (H)ABR. Bij de BGB zijn deze gegevens nodig voor het gaafmaken, bij Baseline worden de gegevens gebruikt om te standaardiseren en bij Impect II om gaaf te maken en te wegen. Deze additionele gegevens noemen we gemakshalve referentiedata. Ze worden tijdens de verwerking als een soort referentie gebruikt. Zij zijn wel als input nodig, maar er wordt geen 'waarde' aan toegevoegd.

[6] Het voorbeeld klopt niet helemaal, is verouderd, maar geeft wel aan hoe processen in de praktijk aangeduid worden. Daarbij gaat niet zozeer om de afkortingen zelf – deze zijn voor buitenstaanders vaak nietszeggend – maar om de boodschap dat iedere afkorting voor een lokale verwerkingsomgeving staat.

Indien we een proces opvatten als een systeem waarbij de inputs aan de linkerkant het systeem binnenkomen, de middelen (referentiedata) aan de onderkant en de outputs aan de rechterkant het systeem uit gaan, dan kan het hierboven gegeven voorbeeld worden weergegeven als in figuur C-2.

Opvallend aan figuur C-2 is dat er in de keten om omzetcijfers te maken meerdere lokale systemen (zeg omgevingen) betrokken zijn en dat deze omgevingen rechtstreeks met elkaar de benodigde statistische data uitwisselen. Figuur C-2 beschrijft daarmee een huidige, misschien enigszins gedateerde, situatie.

De Business Architectuur van het CBS schrijft voor dat in de toekomstige 'ideale' situatie de (interne) uitwisseling van data niet rechtstreeks moet verlopen, maar via koppelvlakken met een standaard uitwisselingsformaat, zie figuur C-3. Bovendien stelt de Business Architectuur van het CBS dat niet alleen de statistische data worden uitgewisseld, maar ten minste ook de bijbehorende metadata. We vatten bovengenoemde eisen vanuit de Business Architectuur samen met de volgende twee principes.

- Er is geen data zonder metadata
- Data en metadata moeten algemeen toegankelijk zijn

Het zijn dus de statistische gegevens die intern worden uitgewisseld en niet louter de statistische data. Door de data en metadata algemeen toegankelijk te maken (via koppelvlakken met standaard uitwisselingsformaten) wordt hergebruik van gegevens bevorderd en wordt de keten efficiënter.

5.2 Rustpunten

Nu is het niet zo dat alle in de lokale omgevingen beschikbare data (intern) moeten worden uitgewisseld en dus via een koppelvlak beschikbaar gesteld. Er gelden vier voorwaarden voor hergebruik, waarvan de eerste twee het gevolg zijn van het principe dat er geen data is zonder metadata.

- De inhoudelijke betekenis van de data is helder gedefinieerd. Willeboordse, Struijs en Renssen (2006) onderkennen verschillende ambitieniveaus
    1. gebruik van 'well defined concepts', dat wil zeggen de populatie, de attribuutvariabelen en de perioden waarvoor de objecten en de eigenschappen van deze objecten geldig zijn, zijn eenduidig gedefinieerd;
    2. gebruik van een 'uniform language', dat wil zeggen, dat concepten met dezelfde inhoudelijke betekenis dezelfde naam dragen, en andersom dat concepten met dezelfde naam dezelfde inhoudelijke betekenis hebben;
    3. het streven naar 'standardization', dat wil zeggen dat bij het ontwerp van de 'well defined concepts' voor een bepaald thema, gebruik wordt gemaakt van een standaard ***CBS-brede lijst*** van 'well defined concepts' voor dit thema;

4. het streven naar ‘harmonization, dat wil zeggen de CBS-brede lijst van ‘well defined concepts’ zo klein en zuiver mogelijk te houden door tussen ‘nearby-concepts’ keuzes te maken welke wel en welke niet in de lijst opgenomen zullen blijven;

- De kwaliteit van de data helder is gedefinieerd en voldoet aan bepaalde standaarden. Indien er een ‘uitwisseling’ bestaat tussen betrouwbaarheid en actualiteit van een dataset, kunnen meerdere kwaliteitsversies worden gedefinieerd (denk aan flashramingen, voorlopige cijfers, nadervoorlopige cijfers en definitieve cijfers), waarbij betrouwbaarheid is ‘afgekocht’ met actualiteit.
- De relevantie van de data voor (intern) hergebruik blijkt uit meerdere (potentiële) gebruikers (afnemers, klanten).
- De dataset heeft een eigenaar (vaak ook leverancier) die verantwoordelijk is voor ontwerp en realisatie.

Indien een kwaliteitsversie van een dataset via een koppelvlak voor (intern) gebruik beschikaar wordt gesteld dan dient het risico dat dit product ‘uit het koppelvlak’ moet worden genomen vanwege achteraf geconstateerde gebreken klein te zijn. Opgeleverde data zijn in rust. Dit in tegenstelling tot data die nog in productie zijn en misschien nog een groot aantal verbeterslagen nodig hebben. Om deze reden en om de reden dat een oplevering van een kwaliteitsversie vaak een proces afsluit, wordt in de CBS-wandelgangen een aan een koppelvlak opgeleverde dataset ook wel een ***rustpunt*** genoemd.

De rustpunten vormen de schakels in de keten. Vanwege de eisen die aan rustpunten worden gesteld, zijn zij een eerste belangrijke stap om transparantie en reproduceer-baarheid in de keten te vergroten.

5.3 Halen of brengen

Idealiter wordt een rustpunt, i.e. een statistische dataset die aan de voorwaarden voor hergebruik voldoet, zodanig ontworpen dat met dit éne product in de behoefte van een zo groot mogelijk aantal gebruikers kan worden voorzien. Over het algemeen betekent dit dat een rustpunt grotere populaties afbakent en meer variabelen bevat dan een individuele gebruiker nodig heeft. Voor het maken van de klantspecifieke selecties geldt het principe

- de eigenaar brengt (naar koppelvlak), de klant haalt (uit koppelvlak) en maakt haar eigen specifieke selecties.

Refererend naar figuur C-3 kan worden gesteld dat de pijlen naar het koppelvlak toe worden uitgevoerd door de eigenaar en de pijlen van het koppelvlak af door de gebruiker. Merk op dat de specifieke klantselecties al een vorm van duiden is.

Bovenstaand principe is een belangrijke stap om de transparantie, efficiëntie en flexibiliteit van de keten te vergroten. Immers, de eigenaar stelt voor al haar gebruikers een zo beperkt mogelijk aantal rustpunten beschikbaar – dit vergroot de efficiëntie en

transparantie – en iedere gebruiker selecteert hieruit datgene wat nodig is. Dat laatste is niet alleen efficiënt, maar stelt de gebruiker in staat om de selecties naar eigen behoefte op flexibele wijze aan te passen.

5.4 Koppelvlakken

Zoals gezegd worden rustpunten (intern) uitgewisseld via koppelvlakken met vaste uitwisselingsformaten. De Business Architectuur van het CBS onderscheid vier soorten koppelvlakken voor statistische gegevens

- Inputbase; de rustpunten hierin betreffen ruwe gegevens die voldoen aan de waarneemvraag van het CBS (zie ook paragraaf 6.1);
- Microbase; de rustpunten hierin betreffen door het CBS bewerkte gegevens (microdata);
- Statbase; de rustpunten hierin betreffen door het CBS bewerkte gegevens (geaggregeerde data);
- Outputbase; de rustpunten hierin betreffen publiceerbare gegevens voor extern gebruik.

Enerzijds kunnen deze koppelvlakken worden opgevat als de CBS-magazijnen met een vaste (technische) interface waarin statistische gegevens als rustpunten zijn opgeslagen (gearchiveerd) en waaruit statistische gegevens kunnen worden ontsloten. Anderzijds geeft het onderscheid in de koppelvlakken een typering van opeenvolgende toestanden van gegevens aan die een steeds hoger niveau in de CBS-waardeketen representeren.

Zo zitten in de inputbase louter statistische gegevens zoals deze ruw, ie. door het CBS statistisch inhoudelijk onbewerkt, het CBS zijn binnenkomen. Wel kunnen voorafgaand aan de inputbase datamanipulaties plaatsvinden, zoals het selecteren van variabelen of het transfomeren van formaat, als er conform de waarneemvraag ‘teveel’ aan data is waargenomen.

In de outputbase zit louter publicabele statistische gegevens zoals ze het CBS kunnen verlaten, dat wil zeggen dat na de outputbase de gegevens niet meer worden bewerkt. Wel kan aan de gegevens nog informatie worden toegevoegd door ze te duiden (analyseren), te presenteren (visualiseren) of te communiceren (uitleggen).

Statistische gegevens in de microbase of statbase zijn door het CBS bewerkt maar zijn nog niet in het eindstadium. Het belangrijkste verschil tussen microdata in de microbase en corresponderende microdata in de outputbase is de mate van statistische beveiliging. Eenzelfde verschil geldt tussen geaggregeerde data in de statbase en corresponderende geaggregeerde data in de outputbase. Alle statistische data in de outputbase is afdoende statistisch beveiligd zodat deze zonder risico op onthulling aan externe afnemers beschikbaar kunnen worden gesteld. Dit betekent overigens niet dat statistische data in de andere koppelvlakken onbeveiligd zouden zijn. Ook hier gelden beveiligingsvoorwaarden (zie Renssen en Camstra, 2007), maar deze kunnen anders worden ingericht.

Vanuit de CBS-waardeketen geredeneerd, is het van belang om bij statistische data een onderscheid te maken tussen microdata en geaggregeerde data. Zonder het heel scherp te willen en kunnen stellen

1. hebben microdata betrekking op real-world objecten die geïdentificeerd kunnen worden (of zouden moeten kunnen worden) met eenheden uit een eenhedenbase, zie paragraaf 15.1.
2. hebben geaggregeerde data betrekking op populaties (klassen), die geïdentificeerd kunnen worden (of zouden moeten kunnen worden) met categorieën uit de classificatiebase, zie paragraaf 15.2.

De real-world populaties (klassen) waar geaggregeerde data naar verwijst zijn dus altijd opgebouwd uit 'kleinere' real-world objecten.

De ratio achter het onderscheid tussen microdata en geaggregeerde data is dat vanaf de binnenkomst op het CBS bij microdata geen dataverlies is opgetreden, terwijl dit bij geaggregeerde data wel het geval is. In het verleden is de microbase dan ook veelvuldig gepresenteerd als de toekomstige schatkamer van het CBS welke het grondmateriaal bevat voor statistische analyse en beschrijvend statistisch onderzoek.

De meerwaarde van geaggregeerde data is dat niet alleen CBS-afnemers, maar vooral ook het CBS zelf beschrijvend statistisch onderzoek uitvoert: zij verwerkt de microdata tot schattingen voor descriptieve maten voor een gehele populatie of delen hieruit. Door de aggregatieslag is er dataverlies opgetreden, maar er is tegelijk specifieke statistische kennis in de vorm van methoden en/of modellen toegevoegd. Het dataverlies enerzijds en de toegevoegde specifieke kennis anderzijds zijn de belangrijkste argumenten om in de CBS-waardeketen de microdata te onderscheiden van de geaggregeerde data.

In Huigen (2006a en 2006b) worden kenmerken van statistische gegevens genoemd op basis waarvan ze aan één van de vier koppelvlakken kunnen worden toebedeeld. De praktijk moet echter leren hoe moeilijk of gemakkelijk dit zal gaan.

5.5 Ketenregie

Eén van de doestellingen van het CBS is vermindering van de administratieve lastendruk. Eén van de oplossingen om deze doelstelling te realiseren is het hergebruik van statistische data. Daartoe is het belangrijk dat iedere statistische dataset die voor hergebruik in aanmerking komt inhoudelijk helder gedefinieerd is, in een zekere mate van stabiele (kwaliteits)toestand verkeert en een verantwoordelijke eigenaar heeft. Dit is het concept rustpunt als schakel in de keten.

Minstens zo belangrijk is dat de rustpunten gemakkelijk toegankelijk zijn, daartoe dient het concept koppelvlak, en dat over de rustpunten *afspraken* worden gemaakt die 'passen' in de keten waar het rustpunt deel van uitmaakt. Denk bijvoorbeeld aan de keten van omzetgegevens uit figuur C-3. Voor het managen van dergelijke afspraken dient het

concept ketenregie. Dit concept is van belang omdat een gehele keten zo sterk is als haar zwakste schakel.

Bij ketenregie maken we een onderscheid tussen

- een happy ketenflow: alle rustpunten (schakels) worden volgens afspraak geleverd, i.e. gebracht naar het desbetreffende koppelvlak.
- een unhappy ketenflow: één of meer rustpunten worden niet volgens afspraak geleverd. De kwaliteit voldoet niet of de levering is te laat.

Het concept van ketenregie is op dit moment nog onvoldoende uitgewerkt om in deze cursus te kunnen behandelen. Wel kunnen we een aantal ingrediënten bespreken die bij ketenregie een rol spelen.

Ketenregie start bij het ontwerpen van de happy ketenflow. Dit ontwerp behelst niet alleen de keten van de rustpunten met kwaliteitseisen (zie paragaaf 5.2), maar ook het tijdschema wanneer welke kwaliteitsversie van welk rustpunt beschikbaar moet komen.

De kwaliteitseisen en het tijdschema dienen vervolgens geoperationaliseerd te worden in één of meer *kwaliteitsindicatoren* met bijbehorende *normen*. Vanuit de Business Architectuur van het CBS dienen deze kwaliteitsindicatoren niet alleen meetbaar en normeerbaar te zijn, maar ook relevant. Immers, het meten van kwaliteitsindicatoren vraagt om extra inspanningen, die de moeite waard moeten zijn voor de vergaarde kwaliteitsinformatie.

Uiteindelijk dient het ketenontwerp van de happy flow, inclusief de geoperationaliseerde kwaliteitsindicatoren en normen, realistisch en relevant te zijn. Realistisch wil zeggen dat alle eigenaren in de keten van mening moeten zijn dat ontworpen happy flow voldoende haalbaar is. Relevant wil zeggen dat de gebruikers in de keten de ontworpen ketenproducten voldoende bruikbaar vinden. De Business Architectuur van het CBS hanteert daarbij het principe

- er zijn klanten (gebruikers) die eten wat de pot (keten) schaft.

Dit principe houdt in dat bij het ketenontwerp niet alle gebruikers geconsulteerd hoeven worden, maar alleen de meest belangrijke.

Het meten van de kwaliteitsindicatoren gebeurt tijdens de uitvoeringsfase van het statistische proces en resulteert in gemeten kwaliteitsindicatoren. Het toetsen van de kwaliteit is het confronteren van de kwaliteitsmetingen met hun norm en resulteert in het nemen van wel of geen actie.

Binnen de happy ketenflow kan het nemen van een actie nog onderdeel van een regulier proces zijn. De kwaliteitsversie van een rustpunt heeft nog niet haar eindstatus bereikt en er vindt gewoon een volgende iteratie binnen het proces plaats. Binnen het tijdschema is daar bovendien nog voldoende ruimte voor.

Echter, zodra er geconstateerd wordt dat niet aan de gemaakte afspraken kan worden voldoen (of is voldaan), dan is de actie procesoverstijgend. In dat geval spreken we van

een unhappy ketenflow. Daarbij kan nog een onderscheid worden gemaakt of de eigenaar de problemen vooraf constateert of dat een gebruiker dit achteraf constateert. In beide gevallen is enige vorm van ketenregie nodig waarin de 'keten' over het probleem wordt geïnformeerd en waarin de 'keten' tot een oplossing komt. De Business Architectuur hanteert daarbij het principe dat de eigenaar verantwoordelijk is voor de geconstateerde tekortkoming en dat alleen de eigenaar herstel kan uitvoeren en opnieuw kan aanleveren.

6. Het statistische proces in fasen

Het statistische proces is van oudsher ingedeeld in drie fasen: input (waarneming), throughput (verwerking en statistische beveiliging) en output (informatieontwikeling en publicatie).

6.1 Input

Grofweg kan worden gesteld dat het doel van de inputfase (waarneming) is om statistische data die bij de externe leveranciers, respondenten of berichtgevers aanwezig zijn in bijvoorbeeld databanken, hoofden en administraties op het CBS te 'verkrijgen'. In het tegenwoordige IT-tijdperk gaat het daarbij niet zozeer om de data fysiek op het CBS te verkrijgen – ze mogen fysiek buiten het CBS blijven – maar om ze voor het CBS toegankelijk te maken.

Datgene wat we op het CBS als statistisch (tussen)product van externe leveranciers en/of respondenten toegankelijk willen hebben, noemen we in deze cursus de waarneemvraag. Zoals zal blijken in paragraaf 6, kan deze vraag afwijken wat het CBS daadwerkelijk als product krijgt. De Business Architectuur van het CBS stelt dat

- het ontwerp van de waarneemvraag outputgericht tot stand moet komen (het is uiteindelijk ten behoeve van een eindproduct waar externe klanten voor zijn) en
- dat de realisatie van de waarneemvraag als rustpunt via een koppelvlak (inputbase) beschikbaar wordt gesteld als daar meerdere (potentiële) interne gebruikers voor zijn.

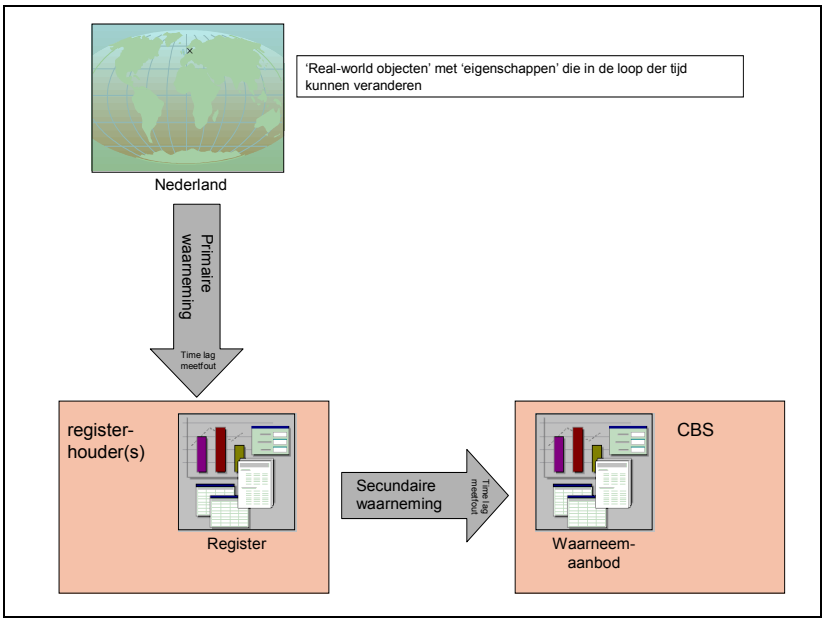
Datgene wat het CBS van registerhouders, respondenten, etc. aan gegevens krijgt noemen we in deze cursus het waarneemaanbod. Bij de realisatie van het waarneemaanbod is een onderscheid tussen primaire en secundaire waarneming:

- Bij primaire waarneming verkrijgt het CBS haar gegevens zonder tussenkomst van een registerhouder.
- Bij secundaire waarneming verkrijgt het CBS haar gegevens met tussenkomst van een registerhouder.

In tegenstelling tot primaire waarneming is niet het CBS eigenaar (en verantwoordelijk voor ontwerp, uitvoering en beheer) van het aanbod, maar de registerhouder. Het CBS

'tapt' als het ware de voor het CBS benodigde data af. Vanuit de optiek van een registerhouder is een register primaire waarneming. Vanuit de CBS-optiek is een waarneemaanbod uit dit register secundaire waarneming, zie figuur C-4.

Figuur C-4: secundaire waarneming



Het 'verschil' tussen waarneemaanbod en waarneemvraag is de operationalisatie van de vraag. In paragraaf 7 wordt hier dieper op in gegaan.

Het beleid van het CBS stelt dat secundaire waarneming de voorkeur verdient boven primaire bij min of meer 'gelijke' kwaliteit. In de praktijk kan dit betekenen dat bij het ontwerp van de waarneemvraag rekening wordt gehouden met secundair beschikbare gegevens.

6.2 Throughput

Grofweg kan worden gesteld dat het doel van de throughputfase (verwerking en statistische beveiliging) is om de waargenomen data te verwerken en te beveiligen tot relevante, betrouwbare en samenhangende statistische gegevens. Daarbij kunnen allerlei tussenresultaten ontstaan die relevant zijn voor intern (her)gebruik.

De Business Architectuur van het CBS stelt dat

- wat we aan statistische gegevens als tussen en/of eindproduct beschikbaar willen stellen outputgericht is ontworpen (aan het eind van de keten is er een klant),
- dat het gerealiseerde eindproduct als rustpunt via een koppelvlak (outputbase) beschikbaar wordt gesteld en
- dat eventuele gerealiseerde tussenproducten als rustpunten via een koppelvlak (microbase of statbase) beschikbaar worden gesteld mits daar interne behoefte aan is (er zijn interne gebruikers).

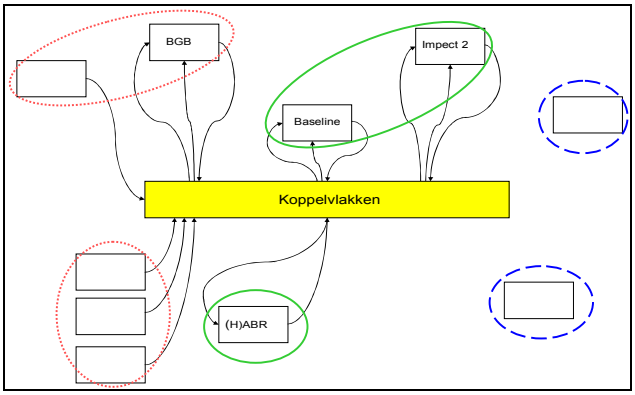
6.3 Output

Grofweg kan worden gesteld dat het doel van de outputfase is om de statistische gegevens als eindproduct te ‘upgraden’ tot statistische informatie en deze informatie voor externe gebruikers te presenteren en toegankelijk te maken. De Business Architectuur van het CBS stelt dat deze statistische informatie gebaseerd is op de verwerkte en beveiligde statistische gegevens in de outputbase en dat zij klantgericht wordt gepresenteerd en toegankelijk gemaakt.

6.4 Van procesfase naar business domein

De drie fasen input, throughput en output hebben onderling sterk afwijkende procesgangen, waarvan elk een eigen bijdrage heeft aan de waardeketen van het CBS. Deze bijdragen zijn via de rustpunten in de koppelvlakken niet alleen helder gedefinieerd, maar ook onderling ‘ontkoppeld’ (niet verweven / autonoom). Om deze redenen zullen we de procesfasen in het vervolg als (business) domeinen aanduiden. In de komende drie paragrafen wordt op elk van de domeinen dieper ingegaan.

Figuur C-5: netwerk tussen verwerkingssystemen (met koppelvlak)



Ter illustratie zijn in figuur C-5 de verschillende procesomgevingen (systemen) om de omzetgegevens te maken gepositioneerd in één van de drie domeinen. De gestippelde kleur rood geeft het inputdomein aan, de doorgetrokken kleur groen het throughputdomein en de streepjes kleur blauw het outputdomein.

7. Het domein waarnemen

Het doel van waarnemen is om externe statistische data voor het CBS toegankelijk te maken. Terugredenerend van een externe informatiebehoefte aan het einde van de keten naar een waarneembehoefte aan het begin van de keten, volgt een concreet ontwerp voor de waar te nemen data. Dit ontwerp noemen we de waarneemvraag. Het bevat de onderdelen

- Conceptuele metadata in een voorschrijvende rol ten aanzien van doelpopulatie (populatietype), doelvariabelen (attribuutvariabelen) en referentieperioden,
- Kwaliteitsmetadata in een voorschrijvende rol ten aanzien van de actualiteit en de betrouwbaarheid.

In termen van de waarneemvraag kan het doel van waarnemen worden geformuleerd als het toegankelijk maken van de bijbehorende data. Daarbij gelden de volgende uitgangspunten.

- Tijdens het ontwerp van de waarneemvraag is vastgesteld dat de benodigde gegevens voor het realiseren van deze vraag niet in de verschillende koppelvlakken (te weten inputbase, microbase, statbase, outputbase) aanwezig zijn.
- Tijdens het ontwerp van de waarneemvraag is vastgesteld dat de benodigde gegevens wel of niet aanwezig zijn in een extern Register, waarbij het Register voor het CBS toegankelijk is conform de CBS-wet.
- Het ontwerp van het externe Register is vastgesteld door de Registerhouder (eventueel in samenspraak met andere partijen waaronder het CBS).
- Het ontwerp van de benodigde gegevens uit het Register, i.e. de waarneemvraag, is vastgesteld door het CBS (eventueel in samenspraak met de registerhouder).
- Een Registerhouder archiveert haar eigen Register.
- Bij geschikt gebleken gegevens in een Register krijgt secundaire waarneming de voorkeur boven primaire waarneming.

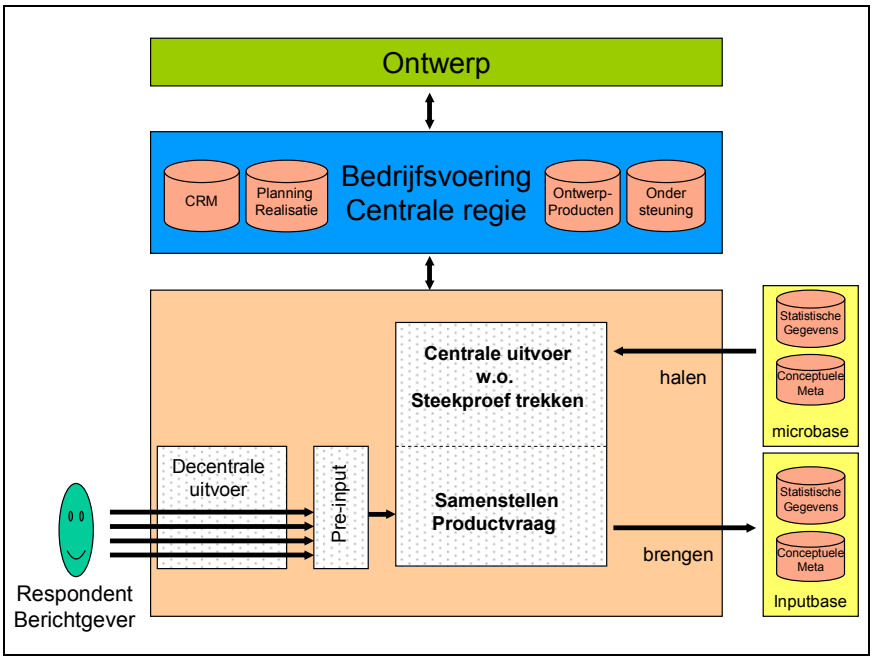
Volgens de uitgangspunten kan dus alleen tot secundaire waarneming worden overgegaan als de benodigde gegevens niet (intern) aanwezig zijn in één van de koppelvlakken, maar wel in een (extern) Register. Tot primaire waarneming kan alleen worden overgegaan als de benodigde gegevens ook niet in een (extern) Register aanwezig zijn.

In deze paragraaf behandelen we eerst het domein primaire waarneming en daarna het domein secundaire waarneming. Dit doen we om didactische redenen, waarbij we af en toe bij de behandeling van secundair waarnemen zullen terugverwijzen naar herkenbare concepten uit de primaire waarneming.

7.1 Primaire waarneming

In figuur C-6 is op hoog abstractieniveau het domein *primaire waarneming* schematisch weergegeven in een aantal subdomeinen. Deze subdomeinen kunnen worden onderverdeeld naar *ontwerp*, *centrale regie* en *uitvoer*. Voor het subdomein *uitvoer* geldt nog weer een verdere opdeling. Er is namelijk één *centraal uitvoerdomein* voor o.a. de steekproeftrekking en er zijn meerder *decentrale uitvoerdomeinen* voor de enquêtering, namelijk voor ieder waarneemkanaal één. Momenteel zijn er vier waarneemkanalen (face-to-face, telefonisch, elektronisch en postaal). Al deze kanalen hebben met elkaar gemeen dat zij het ‘contactenverkeer’ verzorgen tussen het CBS en haar respondenten.

Figuur C-6: het domein *primaire waarneming* in hoofdlijnen



Voor de realisatie van de primair uit te vragen waarneemvraag, dient de vraag te worden geoperationaliseerd. Daarbij gaat het om de volgende stappen.

- Van doelpopulatie (en kwaliteitseisen ten aanzien van betrouwbaarheid) naar waar te nemen populatie en van waar te nemen populatie via populatie- en steekproefkader naar steekproef.
- Van doelvariabelen (en kwaliteitseisen ten aanzien betrouwbaarheid) via vragenlijstontwerp naar vragenlijstsjabloon.
- Van kwaliteitseisen ten aanzien van actualiteit en betrouwbaarheid naar procesontwerp voor enquêtering, inclusief kwaliteitsindicatoren en bijbehorende normen.

Merk op dat het subdomein *centrale waarneemregie* stuurt op de kwaliteitsindicatoren en bij tekortkomingen *ketenregie* informeert en consulteert over noodzakelijke acties die het primaire waarneemdomein overstijgen.

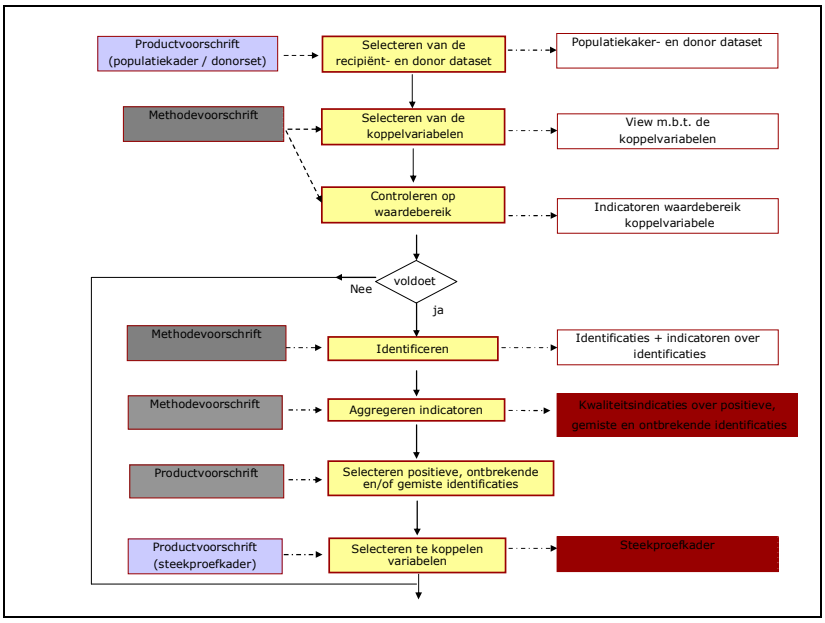
7.1.1 Van doelpopulatie naar steekproef

In de praktijk kan het lastig zijn – vooral bij bedrijfsstatistieken – om de complete set van gewenste gegevens over één doelobject bij één respondent/berichtgever uit te vragen. Om deze reden wordt het doelobject ‘gesplitst’ in één of meer waarneemobjecten waarover de gewenste gegevens wel bij één respondent/berichtgever uit te vragen zijn. De doelpopulatie is de populatie van real-world doelobjecten. De hieruit afgeleide real-world waar te nemen objecten vormen de waarnemingspopulatie. De splitsingen moeten overigens later wel weer worden gebundeld.

Figuur D-7 in paragraaf 15.1 laat de afbeelding zien van een populatie van real-world objecten op een statistische dataset (populatiekader). Het maken van deze afbeelding valt buiten het domein waarnemen, maar dit domein kan het populatiekader wel goed als input gebruiken. Daarbij wordt het populatiekader via de microbase aan het primaire waarnemingsdomein beschikbaar gesteld.

Idealiter dekt het populatiekader zowel de waarnemingspopulatie als doelpopulatie (en de relatie daartussen). Het waarnemingsgedeelte kan gebruikt worden als input voor een steekproefkader en het populatiegedeelte voor de bundeling van de waarnemings-eenheden tot de doeleenheden.

Figuur C-7: Standaard processtap *koppelen*



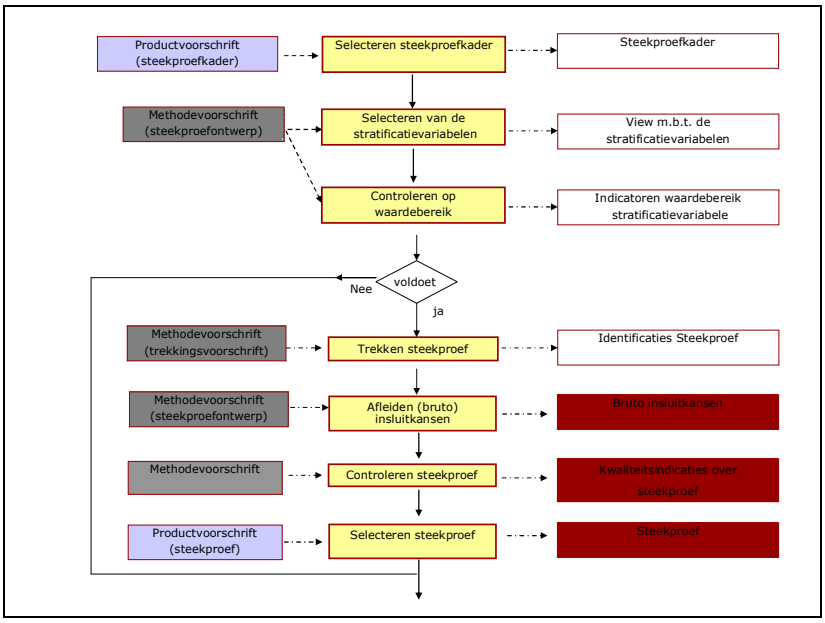
In figuur C-7 is ter illustratie een processtap *koppelen* geschetst om de over- en onderdekking van het populatiekader ten opzichte van de doelpopulatie te meten, en het populatiekader eventueel af te bakenen en/of te verrijken met extra variabelen voor een efficiënt steekproefkader.

In figuur C-8 is ter illustratie een processtap *steekproef trekken* geschetst om uit dit steekproefkader een steekproef te trekken.

Beide processtappen worden uitgevoerd in het domein centrale uitvoer, maar ontworpen in het domein ontwerp en beide kunnen worden opgevat als een standaard processtap (zie paragraaf 9).

Het doel van *koppelen* is om (verrijkt) steekproefkader te verkrijgen, terwijl het doel van *steekproef trekken* is om een steekproef te verkrijgen. Zowel het verrijkte steekproefkader als de steekproef zijn micro datasets met een productontwerp in termen van voorschrijvende conceptuele metadata en voorschrijvende kwaliteitsmetadata. In de figuren 10 en 11 wordt vanwege het voorschrijvende karakter het productontwerp met productvoorschrift aangeduid.

Figuur C-8: Standaard processtap *steekproef trekken*



Merk daarbij op dat in een hiërarchie van doelen en middelen het verrijkte steekproefkader weer dient voor het trekken van de steekproef. Op basis van beide standaard procestappen kan, vooruitlopend op de terminologie van paragraaf 9, een standaard proces *steekproef trekken* worden ingericht. Dit standaard proces bevat dus niet alleen de standaard processtap *steekproef trekken*, maar ter voorbereiding daarop tevens de standaard processtap *koppelen*.

In de wetenschappelijke literatuur bestaan er zeer veel statistische methoden, zeg typen steekproefontwerpen, om een steekproef te trekken. Denk daarbij aan enkelvoudige steekproefontwerpen, gestratificeerde steekproefontwerpen, twee-traps steekproef-ontwerpen, twee-fasen steekproefontwerpen, systematische steekproefontwerpen, etc. Voor een concrete toepassing in het standaard proces *steekproef trekken* wordt een keuze uit zo'n type ontwerp gemaakt – of er wordt een combinatie van typen ontwerpen gekozen – en wordt het gekozen type ontwerp nader gespecificeerd. Deze nadere specificatie wordt in de cursus het steekproefontwerp genoemd.

In het steekproefontwerp wordt de waar te nemen populatie vertaald naar een productvoorschrift voor een steekproefkader en worden de kansen vastgelegd waarmee de steekproef uit dit kader getrokken moet worden.

Het trekken van een steekproef conform een steekproefontwerp kan nog op verschillende manieren. Voor het daadwerkelijk trekken van de steekproef wordt het steekproefontwerp doorvertaald naar een operationeel trekkingsvoorschrift, waarbij wel of geen rekening kan worden gehouden met de zogenaamde enquêtedrukspreiding.

7.1.2 Van doelvariabele naar vragenlijstsjabloon

In de praktijk kan het niet alleen lastig zijn om de complete set van gewenste gegevens over één doelobject bij één respondent/berichtgever uit te vragen, maar de complete set van doelvariabelen is vaak ook niet zonder meer uitvraagbaar.

In het vragenlijstontwerp worden de doelvariabelen met de begripsomschrijvingen geoperationaliseerd, waarbij per waarneemkanaal de operationalisatie kan verschillen. Hier worden onder meer keuzes gemaakt ten aanzien van vraagformuleringen, antwoordbereiken, toelichtingen, volgordes waarin de vragen gesteld worden, optionele vraagstellingen voor specifieke doelgroepen en de presentatievorm (layout).

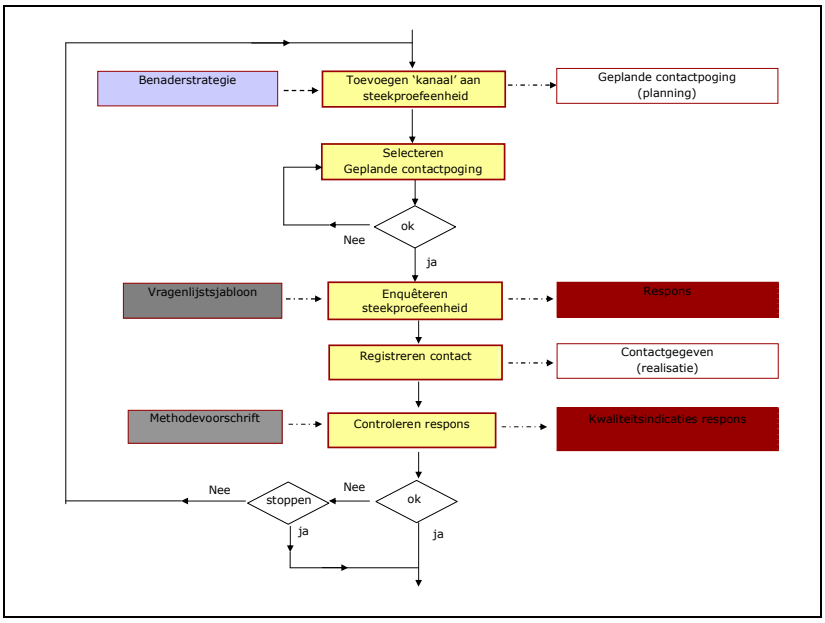
Vanuit het vragenlijstontwerp wordt het (elektronische) vragenlijstsjabloon ontworpen en eventueel worden vanuit het sjabloon papieren vragenlijsten gedrukt. Het zijn de elektronische vragenlijstsjablonen en/of de papieren vragenlijsten die via het betreffende kanaal aan de respondenten worden voorgelegd.

7.1.3 Van kwaliteitseis naar procesontwerp voor enquêtering

Er blijft nog één belangrijke operationaliseringvraag over, namelijk de doorvertaling van de kwaliteitseisen ten aanzien van actualiteit en betrouwbaarheid naar een procesontwerp voor de decentrale uitvoer met centrale regie, zie figuur C-9. Bij dit procesontwerp horen

- een planning om het proces te kunnen starten,
- een benaderstrategie voor de toekenning van een kanaal aan een steekproefeenheid voor een contactpoging (op basis van contactgegevens),
- kwaliteitsindicatoren om de kwaliteit van de respons te meten (monitoren),
- bijbehorende normen om het proces te kunnen beëindigen (of bijsturen).

Figuur C-9: Standaard processtap *enquêteren*



Het enquêteringproces start door aan iedere steekproefeenheid een waarneemkanaal en een planning toe te wijzen. De toewijzing gebeurt op basis van een benaderstrategie en wordt uitgevoerd in het domein centrale regie. Het resulteert in een geplande contactpoging. In het domein decentrale uitvoer worden vervolgens de geplande contactpogingen geselecteerd en zullen de contactpogingen daadwerkelijk worden uitgevoerd conform het toegewezen kanaal. Het resultaat daarvan is in ieder geval een contactgegeven, en bij een respons tevens een ingevulde vragenlijst. Het domein centrale regie controleert het contactgegeven (de realisatie van de planning). Bij een 'mislukt' contact wordt aan de steekproefeenheid opnieuw een waarneemkanaal toegewezen. Bij een 'succesvol' contact wordt de kwaliteit van de respons gemeten. Op basis van de responsmetingen (en een norm) wordt besloten het enquêteringsproces te beëindigen of niet.

Merk op dat bij de decentrale uitvoer niet noodzakelijkerwijs de waarnemingseenheden worden benaderd. Er kan bijvoorbeeld sprake zijn van een afspraak tussen de waarnemingseenheid en het CBS dat de gewenste informatie door een derde partij geleverd wordt. De eenheid die daadwerkelijk wordt benaderd voor een contact om de gewenste gegevens te verkrijgen is de benadereenheid.

Merk verder op dat het ontwerp van de steekproef, de vragenlijst en de benaderstrategie vaak simultaan tot stand komt, omdat het, gegeven een bepaald waarneembudget en kostenmodel, communicerende vaten zijn met betrekking tot het halen van kwaliteitseisen ten aanzien van de betrouwbaarheid (dekking, respons en meetfout).

7.1.4 Van respons naar waarneemvraag

Zoals we het (primair) waar te nemen product hebben opgevat als een waarneemvraag, zo zullen we de respons opvatten als een waarneemaanbod. De respons is namelijk datgene wat door de (externe) respondent is geleverd. We maken dit onderscheid zo expliciet, omdat het ontwerp van de respons kan afwijken van het ontwerp van het gevraagde.

- De waarneempopulatie van de respons verschilt met de doelpopulatie van de vraag.
- De vragen in het vragenlijstsjabloon komen niet één-op-één overeen met de gevraagde doelvariabelen.

De vraag is of de afstemming van het aanbod op de vraag, i.e. de terugvertaling van de operationalisatie, nog tot het domein primaire waarneming behoort. In de cursus gaan we ervan uit dat dit het geval is en dat op basis van de respons de waarneemvraag moet worden samengesteld.

Samenvattend kan daarmee worden gesteld dat het doel van het primaire waarnemingsproces is om aan het ontwerp van de waarneemvraag te voldoen, terwijl de output van het primaire waarnemingsproces de op de respons gebaseerde realisatie van dit ontwerp is, waarbij er methodologische keuzes zijn gemaakt ten aanzien van steekproefontwerp, vragenlijstontwerp en de benaderstrategie. Mits er meerdere gebruikers zijn, wordt de realisatie via de inputbase centraal ter beschikking gesteld.

7.2 Terminologie bij secundaire waarneming

7.2.1 Registerbase

Hoewel het ontwerp van een Register door de Registerhouder wordt gedaan, gaan we in deze paragaaf uit van een prototypisch register, de zogenaamde Registerbase. Een Registerbase bevat één of meer afbeeldingen van Nederland op een logisch datamodel. Het logische datamodel wordt gekenschetst door

- Eén of meer objecttypen en populatieafbakeningen,
- Eén of meer sleutel- en attribuutvariabelen per objecttype
- Referentieperioden voor de waarden van de sleutel- en attribuutvariabelen.

Idealiter is een feit in Nederland foutloos en zonder time lag (vertraging) geregistreerd. Dat betekent onder meer dat alle reële mutaties in real-time en foutloos in de Registerbase terecht komen. In werkelijkheid zal een registratie een vertraging hebben en niet foutloos zijn. In ons prototypische Registerbase wordt daartoe per geregistreerd object een 'dossier' bijgehouden met een 'registratiehistorie'. In deze historie zijn de voor dit object geregistreerde feiten, i.e. de waarden van de sleutel- en attribuutvariabelen met betrekking tot een zekere referentieperiode, tevens aangegeven

- De foutcorrecties
- De registratiemomenten van zowel de originele registraties als de foutcorrecties hierop.

Merk op dat aan de hand van de registratiemomenten de geldigheidperiode van een geregistreerd feit kan worden afgeleid, zie ook paragaaf 3.4.

Refererend naar figuur C-4 kan een Registerbase worden opgevat als soort film over Nederland ten aanzien van één of meer aspecten. Immers, iedere keer als er zich ten aanzien van deze aspecten een mutatie voordoet, dan wordt deze mutatie zo snel mogelijk verwerkt in de Registerbase.

7.2.2 Registerstand

Een Registerstand wordt in deze cursus opgevat als een soort foto (momentopname) over Nederland ten aanzien van één of meer aspecten. Het logische datamodel van een Registerstand wordt gekenschetst door

- Eén of meer objecttypen en populatieafbakeningen
- Eén of meer sleutel- en attribuutvariabelen per objecttype
- Eén referentieperiode voor de waarden van de sleutel- en attribuutvariabelen.

Idealiter, indien de registratie foutloos is, kan een Registerstand eenvoudig worden verkregen uit een Registerbase door op basis van een aangegeven *peilmoment* (deze komt

overeen met de referentieperiode), *populatieafbakening* en *lijst van sleutel- en attribuut-variabelen* de geldige objecten met bijbehorende waarden van variabelen te selecteren.

In werkelijkheid zal de Registerbase fouten bevatten en dient het selectiecriterium uitgebreid te worden met het Registratiemoment. Een Registerstand wordt dan niet alleen door het logische datamodel gekenschetst, maar ook door een registratiemoment.

Onder de aanname dat een foutcorrectie altijd een verbetering inhoudt, is het registratiemoment een soort kwaliteitscriterium. Immers, naarmate het registratiemoment actueler is, zal de Registerstand minder fouten bevatten. Deze kwalitatief verschillende Registerstanden vervullen daarmee een overeenkomstige rol als de verschillende kwaliteitsversies van een rustpunt, zie paragraaf 5.2.

7.2.3 Registervraag

Op basis van enerzijds de statistische informatiebehoefte aan het einde van de keten – deze is dus outputgericht – en anderzijds de in een Registerbase aanwezige gegevens, dien vastgesteld te worden welke gegevens uit de Registerbase onttrokken moeten worden. Dit is het ontwerp van de waarneemvraag bij secundaire waarneming. Bij dit ontwerp onderscheiden we de volgende twee smaken.

- Er is alleen behoefte aan geregistreerde momentopnamen en dus naar een Registerstand, eventueel naar verschillende Registratiemomenten.
- Er is behoefte aan een meest actueel geregistreerde film en dus aan een Registerbase met een ‘compleet’ dossier met een registratiehistorie.

Bij de behoefte aan een Registerstand dienen de gewenste populatieafbakeningen, variabelen, referentieperioden en kwaliteitsversies nader gespecificeerd te worden. Bij de behoefte aan een Registerbase dienen vooral de gewenste objecttypen en variabelen nader gespecificeerd te worden. In de volgende paragaaf zal op beide behoeften nader worden ingegaan.

7.3 Procesmodel voor secundaire waarneming

Deze paragraaf start met de beschrijving van het procesmodel voor het waarnemen van een Registerstand. Daarbij onderscheiden we vier soorten ontwerpen

- Het ontwerp van een Registervraag (waarneemvraag), in dit geval een Registerstand,
- Het ontwerp van een Registeraanbod (waarneemaanbod),
- Het ontwerp van de operationalisatie van de kwaliteit,
- Het ontwerp van de procesflow.

Aan het einde van de paragraaf zal op het procesmodel voor het waarnemen van een Registerbase worden ingegaan.

7.3.1 Registervraag versus Registeraanbod

Een Registervraag is datgene wat het CBS uit een Registerbase bij de Registerhouder wil hebben. Deze vraag is tot stand gekomen vanuit een (externe) informatiebehoefte aan het einde van de keten. De Registervraag komt overeen met de aan het begin van deze paragraaf beschreven waarneemvraag, maar dan specifiek voor secundaire waarneming. Een Registeraanbod is datgene wat door de (externe) Registerhouder wordt geleverd. Idealiter willen we graag dat

- Het ontwerp van een Registervraag outputgericht is en dat het ontwerp van het Registeraanbod hierop is aangepast, maar in de praktijk geldt vaak dat
- Het ontwerp van een Registervraag veelal inputgericht is – geef maar alles – en gelijk gesteld aan het ontwerp van het Registeraanbod, omdat de informatiebehoefte aan het einde van de keten vaak niet op voorhand duidelijk is en ook weer afhangt van de gegevens in de Registerbase bij de Registerhouder.

Om deze schijnbare tegenstelling te overbruggen gaan we in deze cursus van het volgende uit. Een Registervraag wordt zo breed mogelijk ontworpen om tegemoet te komen aan zoveel mogelijk afnemers verderop in de keten. Uit het ontwerp van een Registervraag resulteert een statistisch product (in dit geval een Registerstand) voor de inputbase, waarbij er twee uiterste opties zijn om dit product te verkrijgen.

- De eerste optie is dat het CBS de Registerstand zelf ‘haalt’ uit de Registerbase bij de Registerhouder en deze via de inputbase op het CBS (intern) ter beschikking stelt. Dat wil zeggen, dit domein ontwikkelt de benodigde selectie- en aggregatie-voorschriften – eventueel in samenwerking met de Registerhouder – en voert de selecties en aggregaties uit op de locatie van de Registerhouder. Het CBS is ‘in charge’ van de selectie- en aggregatievoorschriften en kan deze naar eigen inzichten aanpassen. Registervraag en Registeraanbod zijn aan elkaar gelijk.
- Bij de tweede optie komt de gevraagde Registerstand in twee fasen tot stand. In de eerste fase wordt een Registeraanbod uit de Registerbase bij de Registerhouder geselecteerd door de Registerhouder en ‘gebracht’ naar het CBS. In de tweede fase wordt op het CBS en door het CBS uit deze selectie de gevraagde Registerstand samengesteld en (intern) beschikbaar gesteld via de inputbase. Registervraag en Registeraanbod zijn niet noodzakelijkerwijs aan elkaar gelijk.

De eerste optie biedt het CBS de mogelijkheid om het ontwerp van de Registervraag flexibel en tegen relatief lage kosten aan te passen. Immers de Registerhouder stelt het CBS in de gelegenheid zelf het Registeraanbod te definiëren, waardoor het CBS deze kan afstemmen op haar eigen vraag. Dit maakt het tevens mogelijk een potentieel Registeraanbod met bijbehorende selectie- en aggregatievoorschriften vooraf al vast te leggen in de catalogus van de inputbase, maar de fysieke uitvoering van die voorschriften uit te stellen en daarmee nog geen overbodige fysieke data naar het CBS te halen.

7.3.2 Operationalisatie kwaliteit

Een waarneemvraag – en ook een Registerstand – kenmerkt zich doordat de populatieafbakeningen, sets van variabelen, en referentieperioden in het ontwerp expliciet gedefinieerd zijn. De operationalisering van deze voorschrijvende conceptuele metadata laat zich vrij rechtstreeks vertalen naar een set van selectie- en aggregatievoorschriften die, afhankelijk van optie 1 of 2, op de Registerbase bij de Registerhouder of op het Registeraanbod (bij het CBS).

Het ontwerp van een waarneemvraag geeft ook kwaliteitseisen aan. Voor de operationalisatie van de kwaliteitseisen is, net als bij primaire waarneming, nog een extra ontwerpslag nodig. We onderscheiden daartoe

- Een operationalisatie voor de time lag, dekking en meetfout, resulterend in selectievoorschriften die op basis van het Registratiemoment de 'meest' geldende/actuele registerdata t.o.v. een bepaald tijdstip kunnen selecteren
- Een steekproefontwerp resulterend in een trekkingsvoorschrift ter aanvulling op hierboven benoemde selectievoorschriften, die de eenhedenselectie steekproefsgewijs kan uitvoeren.

Merk op dat op deze manier meerdere kwaliteitsversies van een Registerstand kunnen worden verkregen; een kwaliteitsversie met betrekking tot Registratiemoment A, B, etc.

7.3.3 Ontwerp van de procesflow

Bij zowel optie 1 als optie 2 moeten de kwaliteitsversies van een Registerstand naar het CBS 'vervoerd' worden. Daarbij spelen drie (logische)[7] ontwerpaspecten een rol.

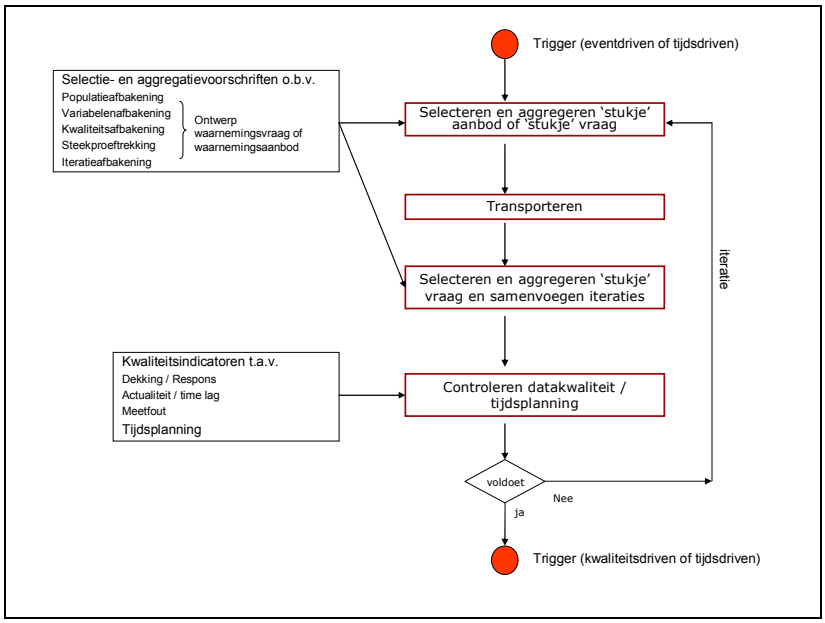
- Vindt de datastroom in één (ononderbroken) transportbeweging plaats of in een iteratieve reeks? In het eerste geval wordt een Registerstand in één keer gebracht of gehaald. In het tweede geval wordt per iteratie een 'stukje' Registerstand gebracht of gehaald. Deze stukjes kunnen overlappend zijn.
- Is de trigger om de transportbeweging (in een eventuele iteratieve reeks) te starten time driven (dus op basis van vooraf gestelde tijdstippen) of event driven (op basis van voorafgestelde gebeurtenissen)? In het eerste geval is het aantal eenheden in een iteratie doorgaans variabel. In het tweede geval is het aantal eenheden in een iteratie doorgaans klein. In het geval dat de overeengekomen gebeurtenis een mutatie op een eenheid betreft, is de omvang van de iteratie zelfs minimaal (precies één).
- Is de trigger om een iteratie reeks van transportbewegingen te beëindigen time driven (vast tijdstippen) of quality driven (een minimale kwaliteit moet worden behaald)? In het eerste geval is de kwaliteit van het product variabel. Vooraf is

[7] De fysieke aspecten vallen buiten de scope van deze cursus.

immers niet bekend welke kwaliteit binnen het gestelde tijdsinterval wordt gerealiseerd. In het tweede geval wordt doorgegaan met het ontvangen of halen van ‘stukjes’ Registerstand totdat de vooraf vastgestelde kwaliteit is behaald.

In figuur C-10 is het proces van secundaire waarneming op een high level schematisch weergegeven. In de figuur is het proces als een iteratieve reeks van transportbewegingen gemodelleerd. Naast de selectie- en aggregatievoorschriften die nodig zijn om uit de Registerbase bij de Registerhouder de gevraagde of aan te bieden Registerstand af te bakenen, zijn additionele selectievoorschriften nodig om de ‘stukjes’ af te bakenen.

Figuur C-10: Secundair waarnemen



Merk op dat het verschil tussen optie 1 en 2 uit paragraaf 7.3.1 vooral zit in waar (op CBS locatie of bij de Registerhouder) en door wie (CBS of Registerhouder) de verschillende activiteiten worden uitgevoerd.

7.3.4 Waarnemen van een Registerbase

Figuur C-10 geeft niet alleen een high level procesmodel weer voor het waarnemen van een Registerstand. Hetzelfde model is ook van toepassing voor het waarnemen van een Registerbase. Ook voor een Registerbase gelden de (uiterste) opties 1 of 2, en juist voor een Registerbase zal het transport iteratief zijn.

De belangrijkste verschillen met het waarnemen van een Registerstand zitten niet zozeer in het de structuur van het proces, maar in het ontwerp van het product en daarmee ook in het ontwerp van de in het proces gehanteerde selectie- en aggregatievoorschriften. We noemen de volgende verschillen.

– Bij het ontwerp van een Registerstand wordt de mogelijkheid open gelaten of dit op basis van een steekproeftrekking is of niet. Dit is afhankelijk van de gevraagde kwaliteit. Het ontwerp van een Registerbase zal in principe ‘integraal’ zijn.

- Verder is bij het ontwerp van een Registerstand de populatieafbakening mede in termen van gesloten tijdsintervallen gedefinieerd (maand, kwartaal, jaar), terwijl de populatieafbakening bij het ontwerp van een Registerbase vanaf een bepaald moment is gedefinieerd.
- Voor de operationalisatie voor de time lag en de meetfout worden bij zowel een Registerbase als een Registerstand de 'meest' actuele/geldige registerdata ten opzichte van een bepaald tijdstip geselecteerd. Echter, bij een Registerstand liggen deze tijdstippen relatief ver uit elkaar, terwijl bij een Registerbase deze tijdstippen 'lopend' zijn.

7.4 Voorbeeld IIS

Het IIS (Inkomens Informatie Systeem) 'verzamelt' zo compleet en actueel als mogelijk inkomens- en bedrijfsgegevens – de zogenaamde IIS-leveringen – en bewerkt deze zodanig voor dat ze geschikt zijn voor het samenstellen van de inkomens- en vermogensstatistieken van Nederland.

Deze IIS-leveringen komen overeen met stukjes Registeraanbod van de Belastingdienst. De Belastingdienst is dus Registerhouder.

- De populatieafbakening van deze leveringen, i.e. stukjes Registeraanbod, zijn alle belastingplichtigen voor Nederland, waarbij de leveringen over meerdere referentieperioden (jaren) kunnen gaan.
- De set van variabelen is zeer groot (meer dan 2.000 variabelen). Sommige variabelen zijn zeer gedetailleerd. Zo worden bijvoorbeeld de detailposten van de giften geleverd, welke een aanslagplichtige in zijn aangifte inkomstenbelasting invult. Voor de Belastingdienst zijn deze posten van belang om te kunnen beoordelen of ze rechtmatig zijn. Voor het CBS zijn deze posten te gedetailleerd.
- De totale stroom aan leveringen dekt in principe de totale populatieafbakening. In principe wordt per belastingplichtige het gehele dossier geleverd, dus alle reële en niet-reële (foutcorrecties) mutaties in dit dossier.

Bovenstaande punten geven een globaal beeld van wat het CBS krijgt van de Belastingdienst. Deze leveringen komen echter niet overeen met wat het CBS wil hebben, i.e. de IIS-leveringen zijn nog niet geschikt voor het samenstellen van de inkomens- en vermogensstatistieken van Nederland. Vandaar dat de IIS-leveringen worden voorbewerkt tot de gewenste IIS-vraag.

- De populatieafbakening van een IIS-vraag zijn alle belastingplichtigen voor Nederland voor jaar xxx.
- De set van variabelen is (wat) kleiner en bovendien zijn sommige variabelen minder gedetailleerd.

- Er zijn meerdere kwaliteitsversies ten aanzien van dekking, actualiteit en meetfout gedefinieerd. Ten aanzien van de nauwkeurigheid geldt dat het steekproefontwerp 'integraal' is.

De voorbewerking bij het IIS houdt in dat uit de min of meer continue stroom van leveringen de gewenste kwaliteitsversies worden samengesteld en opgeslagen in de inputbase.

8. Het domein verwerken en statistische beveiliging

Het doel van het domein verwerken is om waargenomen statistische data – de waarneemvraag – te verwerken en te beveiligen tot relevante, samenhangende en betrouwbare statistische gegevens. Binnen het verwerkingsdomein onderscheiden we vier soorten statistische gegevens, zodat binnen dit doel nog een typering kan worden aangebracht

- Microdata: het verkrijgen van relevante, samenhangende en betrouwbare microdata inclusief kwaliteitsinformatie[8] over de relevantie, samenhang en betrouwbaarheid van de microdata,
- Geaggregeerde data: het verkrijgen van relevante, samenhangende en betrouwbare schattingen voor populatieparameters, inclusief kwaliteitsinformatie[4] over de relevantie, samenhang en betrouwbaarheid van deze schattingen,
- Geïntegreerde data: het verkrijgen van relevante, samenhangende en betrouwbare schattingen voor populatieparameters, die bovendien consistent zijn met restricties die gesteld zijn vanuit integratiestelsels, inclusief kwaliteitsinformatie[4] over de relevantie, (in)consistentie en betrouwbaarheid van de 'ingepaste' schattingen,
- Tijdreeksen: het verkrijgen van relevante, samenhangende en betrouwbare seizoensgecorrigeerde tijdreeksen, inclusief kwaliteitsinformatie[4] over de relevantie, samenhang en betrouwbaarheid van deze tijdreeksen.

Deze typische einddoelen betreffen statistische gegevens waarvoor externe afnemers bestaan. Naast deze typische einddoelen onderscheiden we ook typische tussenliggende doelen. Tussenliggende doelen betreffen statistische gegevens waarvoor (louter) interne afnemers bestaan. Deze interne statistische gegevens betreffen ook microdata, schattingen voor populatieparameters en tijdreeksen, maar voor extern gebruik zijn zij niet relevant, niet samenhangend, of niet betrouwbaar genoeg.

Een set van statistische gegevens dat als tussen of einddoel wordt ontworpen, komt overeen met het in paragaaf 5.2 geïntroduceerde begrip *rustpunt*. Refererend naar figuur

[8] Dit is inclusief informatie over de ontstaansgeschiedenis, voorzover deze geschiedenis nodig is om de data op kwaliteit te kunnen beoordelen.

C-3 heeft ieder verwerkingsproces haar eigen set van gedefinieerde rustpunten en dus haar eigen concreet gemaakte doelen die ze moet realiseren.

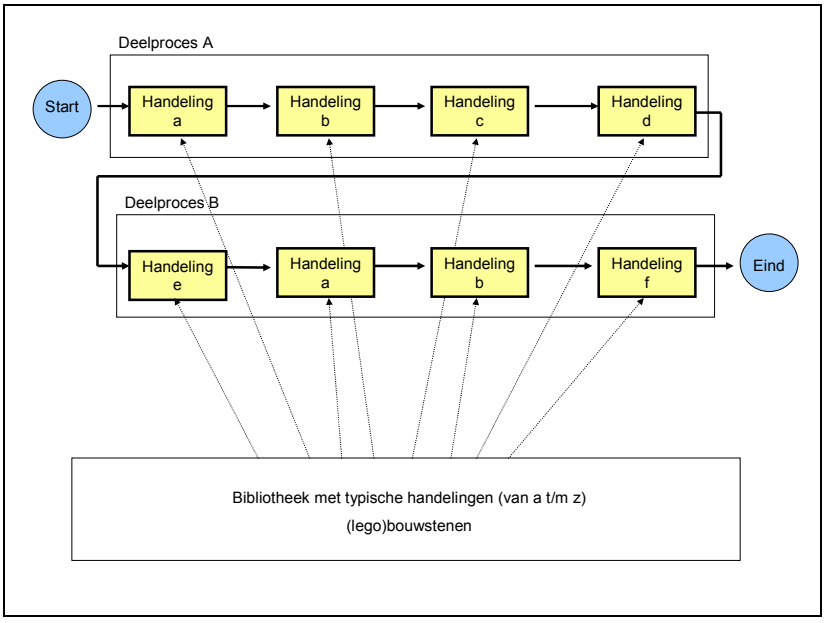
8.1 Over processen en handelingen

Stel dat de statisticus niet een statisticus was, maar een fietsenmaker die een band moet plakken. Typische handelingen daarbij zijn *fiets omdraaien (of fiets ophangen), ventiel losdraaien, buitenband lichten en binnenband vrijmaken, water halen, binnenband door water halen en gaatje merken, rondom gaatje drogen en schuren, plakkertje pakken, knippen en plakken, binnen- en buitenband controleren, steentje verwijderen*[9]*, binnen- en buitenband terugzetten, fiets terugdraaien, ventiel terugdraaien* en *band oppompen*.

Voor een aantal van deze typische handelingen heeft de fietsenmaker een mooi stuk gereedschap. Voor het lichten van de buitenband heeft hij vaak drie bandenlichters beschikbaar en voor het zoeken van het gaatje een emmer (met water).

Een dergelijk samenhangend geheel van handelingen, welke is getriggerd naar aanleiding van een plan of een toestand, met een beoogde output en met verschillende inputs, noemen wij een *proces*. Indien de handelingen gericht zijn op het verwerken van statistische data en de beoogde output dan ook statistische data betreft, dan spreken wij van een *statistisch verwerkingsproces*.

Figuur C-11: een bibliotheek van handelingen



Het fietsenmakersproces *band plakken* heeft als beoogde output een geplakte band en is getriggerd door het hebben van een lekke band. Dit proces van *band plakken* willen we veralgemeniseren naar andere processen, in het bijzonder statistische verwerkings-processen. Figuur C-11 geeft een schematische weergave van een sterk vereenvoudigd

[9] Steentje is ter illustratie. Algemener mag het een USO zijn (Unidentified Sharp Object).

verwerkingsproces. Er is een hoofdproces, welke is opgebouwd uit twee deelprocessen, waarbij ieder deelproces weer is opgebouwd uit handelingen. Daartoe is onderaan de figuur een bibliotheek van typische handelingen weergegeven, waarbij het zeer goed mogelijk is dat verschillende deelprocessen (deels) gebruik maken van dezelfde handelingen.

Enigszins vergelijkbaar met de handelingen van een fietsenmaker wordt in deze nota onder een handeling van een statisticus verstaan een *op zichzelf staande verrichting* met een *eigen functionaliteit*.

- Met 'op zichzelf staand' wordt bedoeld dat de handeling autonoom is, in die zin dat helder is wat de aard van de inputs en randvoorwaarden zijn, wat de aard van de outputs zijn, en dat de handeling contextonafhankelijk is.

- Met 'eigen functionaliteit' wordt bedoeld dat het helder is wat de handeling doet, i.e. voor welke doelen de handeling ingezet kan worden.

Voorbeelden van typische handelingen bij het verwerken van data zijn

- Selecteren / filteren van eenheden
- Selecteren van variabelen
- Hercoderen van variabelen
- Afleiden van variabelen
- Aggregeren over eenheden
- Samenvoegen van bestanden
- Sorteren

Merk op dat deze handelingen vaak al als 'voorgeprogrammeerde' modules in statistische programmatuur beschikbaar zijn. Juist vanwege de autonomie zijn handelingen geschikt om als componenten te worden gebruikt om processen in te richten.

Handelingen zijn de 'kleinste' bouwstenen die we in deze nota onderscheiden bij het modelleren en standaardiseren van statistische verwerkingsprocessen. Net als bij het bandenplakvoorbeeld zijn de handelingen van een statisticus op een min of meer 'natuurlijke' manier ingegeven en geven zij een beschrijving van een verwerkingsproces. In principe zouden handelingen ook weer verder ontleed kunnen worden in nog weer kleinere brokstukjes. Dit leidt vaak tot een meer fysieke beschrijving van een proces en is vaak afhankelijk van het gereedschap dat wordt ingezet om de handelingen te verrichten (bijvoorbeeld band oppompen bestaat uit de hendel van een fietspomp op en neer bewegen).

Voor een goed begrip is het ten slotte goed om een onderscheid te maken tussen

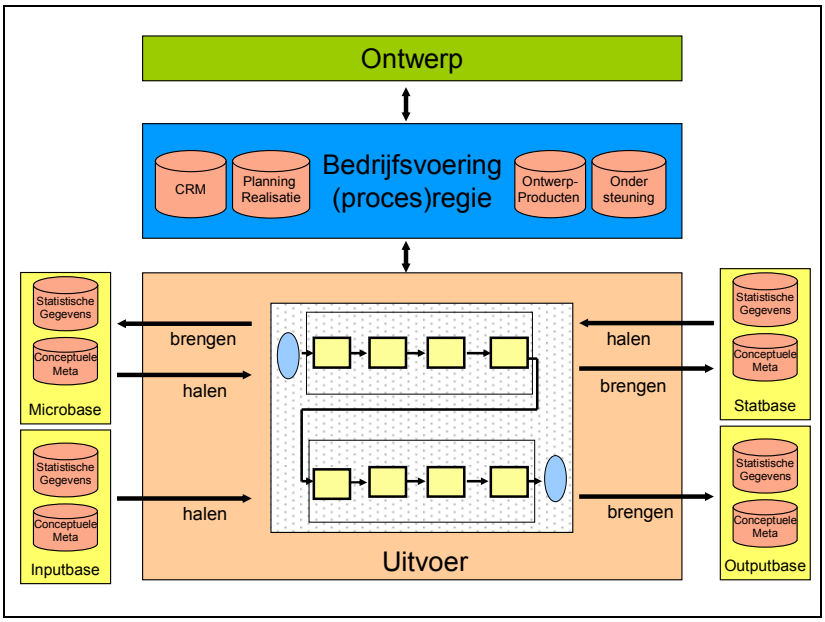
- Typische handelingen zoals ze als een algemene contextloze activiteit kunnen worden omschreven
- Specificaties van typische handelingen zoals ze in een concreet verwerkingsproces zijn verbijzonderd.

Een voorbeeld van een typische handeling is het *selecteren van eenheden*, terwijl een voorbeeld van een specificatie hiervan is het *selecteren van personen ouder dan 15 jaar*.

8.2 Het verwerkingsdomein in hoofdlijnen

Figuur C-12 geeft op hoog abstractie niveau het domein *verwerken* schematisch weer. Evenals bij het domein *waarnemen* is het domein *verwerken* onderverdeeld in drie subdomeinen, namelijk ontwerp, (proces)regie en uitvoer.

Figuur C-12: Het verwerkingsdomein in hoofdlijnen



In het subdomein *ontwerp* worden de te realiseren rustpunten en de daarvoor benodigde methodologische oplossingen en verwerkingsprocessen ontworpen. Uiteraard voldoet het ontwerp van een rustpunt aan de in paragraaf 52 gestelde voorwaarden.

In het subdomein *uitvoer* worden de ontworpen rustpunten gerealiseerd conform de ontworpen verwerkingsprocessen. Ter illustratie is in figuur C-12 het sterk vereenvoudigde verwerkingsproces uit figuur C-11 overgenomen.

In het subdomein *(proces)regie* wordt het uitvoerende proces en de daaruit resulterende rustpunten op voortgang en kwaliteit gemonitord conform de planning en de ontworpen proces- en kwaliteitsindicatoren. Zonodig wordt het proces bijgestuurd en bij procesoverstijgende tekortkomingen wordt *ketenregie* geïnformeerd en geconsulteerd.

De Business Architectuur schrijft voor dat gerealiseerde rustpunten worden gebracht en benodigde rustpunten worden gehaald. Afnemers die selecties nodig hebben uit de gebrachte rustpunten – en deze conform de architectuur zelf moeten halen – kunnen eventueel ondersteund worden.

In paragraaf 9 wordt het verwerkingsproces verder uitgewerkt. Daarbij wordt tevens de relatie met methodologie aangebracht.

8.3 Doel versus output van een verwerkingsproces

Er is een subtiel verschil tussen het doel van een verwerkingsproces en haar output. Dit verschil kan aan de hand van de geïntroduceerde terminologie als volgt worden verwoord:

- Een doel van een verwerkingsproces is een in het ontwerpdomein gedefinieerde kwaliteitsversie van een rustpunt in termen van afgebakende populaties, sets van variabelen en kwaliteitseisen (dit is louter de wat-vraag).
- Een output van een verwerkingsproces is een in het uitvoeringsdomein gerealiseerde kwaliteitsversie van een rustpunt, waarbij de realisatie is gestuurd vanuit een in het ontwerpdomein ontworpen methodologische oplossing welke in het verwerkingsproces is geïmplementeerd (hier zit ook de hoe-vraag in).

Dit inzicht is niet alleen van belang bij het beschrijven van verwerkingsprocessen, maar biedt ook mogelijkheden om de verwerkingsprocessen te structureren en standaardiseren. Verwerkingsprocessen verschillen namelijk vooral in eigen gekozen methodologische oplossingen. Wordt van de oplossing geabstraheerd, dan komen de doelen bovendrijven, die vaak wel hetzelfde zijn en waarvoor generieke methodologische oplossingen bestaan.

9. Contouren van het verwerkingsproces

In figuur C-3 is ter illustratie een aantal systemen schematisch weergegeven voor het verwerken van data. CBS-breed zijn er in de huidige situatie honderden van dergelijke systemen. Veel van deze systemen bestaan uit een grote verzameling van handelingen die weinig transparant is. Figuur C-3 geeft dan ook voor de meeste huidige processen een te gestileerd beeld. Door de intransparantie zijn veel systemen en daarmee de verwerkingsprocessen die ze vertegenwoordigen duur in onderhoud en weinig flexibel.

Vanuit de Business Architectuur van het CBS worden voorwaarden gesteld aan het ontwerp een verwerkingsproces:

- effectief en tijdig: het proces doet wat het moet doen, i.e. de output van het proces is conform het beoogde doel en planning.
- efficiënt en flexibel: gegeven het beoogde doel en planning wordt de output aan de hand van een goedkope methodologische oplossing verkregen. Bovendien kan de oplossing eenvoudig worden aangepast als daar aanleiding toe is.
- transparant en reproduceerbaar: het proces is geen black box en kan als het moet eenvoudig herhaald worden waarbij dezelfde output wordt verkregen.

In paragraaf 5 is een belangrijke stap gezet om de transparantie en reproduceerbaarheid van een keten te verbeteren. In deze paragraaf zal een belangrijke stap worden gezet om de afzonderlijke verwerkingsprocessen te verbeteren. Kernbegrippen bij deze verbeterslag zijn standaard processtap en standaard proces.

9.1 Standaard processtap

Zoals gezegd heeft iedere handeling in een proces een eigen functionaliteit en al deze functionaliteiten zullen bijdragen tot de beoogde output van het proces. Bij het vaststellen van de beoogde output en de daarbij behorende handelingen is echter impliciet dan wel expliciet de keuze gemaakt dat deze output de oplossing is voor een beoogd doel.

Hoewel de beoogde output van het proces *band plakken* een *geplakte band* is, is het beoogde doel van dit proces niet om *een geplakte band te hebben*, maar om *een gave (luchtdichte) band te hebben*. De oplossing was om de band te plakken, wat de reeks van handelingen van *fiets omdraaien* tot en met *band oppompen* tot gevolg had. Evengoed had de oplossing kunnen zijn om de band te vervangen. Dit had geleid tot een andere reeks van handelingen (met hier daar wel dezelfde handeling) en dus tot een ander proces met de output *een nieuwe band*. Dit is een andere output voor hetzelfde doel: *een gave (luchtdichte) band*.

Het expliciet onderkennen van doelen in het verwerkingsproces en het afbakenen en classificeren van de verwerkingsprocessen naar deze doelen is een eerste stap om verwerkingsprocessen transparant te maken. We weten dan waarvoor het proces dient.

Het expliciet onderkennen van het beoogde doel van een reeks van handelingen in de context van een proces is dan ook de grondgedachte van een standaard processtap. Een standaard processtap is evenals een handeling een op zichzelf staande verrichting met een bepaalde functionaliteit. Echter, de functionaliteit van een standaard processtap is statistisch van aard. Dat wil zeggen dat de functionaliteit van een standaard processtap geïnspireerd is vanuit een statistisch doel met een methodologische oplossing daarvoor. Vanwege de veelal complexere functionaliteit is een standaard processtap vaak opgebouwd uit meer dan één handeling.

- Een statistisch doel in het verwerkingsdomein is het bewerken van statistische gegevens, waardoor de relevantie ervan wordt vergroot of de samenhang dan wel de betrouwbaarheid wordt verhoogd. Statistische doelen vinden hun weerslag in ontworpen kwaliteitsversies van rustpunten, zie ook paragaaf 8.3.
- Een methodologische oplossing is een systematische oplossing welke is gebaseerd op één of meer (statistische) methoden en/of op inhoudelijke materiekennis. Methodologische oplossingen vinden hun weerslag in inhoudelijke voorschriften, zie paragraaf 9.4.

Voor een goed begrip is het van belang om onderscheid te maken tussen een generieke oplossing en een specificatie van een generieke oplossing. Een voorbeeld van een generieke oplossing om populatietotalen te schatten is *wegen met een regressiemethode*. Een voorbeeld van een specificatie hiervan is *wegen naar leeftijd en geslacht met een regressiemethode.*

Tevens is het van belang om een onderscheid te maken tussen generieke doelen en specificaties hiervan. Een voorbeeld van een generiek doel is *het verkrijgen van een relevante en betrouwbare microdataset*. Een voorbeeld van een specificatie hiervan is een

ontworpen rustpunt (microdataset) in termen van populatieafbakening(en), set(s) van variabelen en kwaliteitseisen.

Er geldt dat er voor een generiek doel meerdere generieke oplossingen kunnen zijn en dus meerdere soorten standaard processtappen. Voor een specifiek doel kunnen er niet alleen meerdere generieke oplossingen zijn (en dus meerdere soorten standaard processtappen), maar daarbinnen weer meerdere specifieke oplossingen. Elk van deze oplossingen resulteert in een andere – en daarmee specifieke – standaard processtap.

Het is tijd om een standaard processtap uit te werken. Dit zullen we doen aan de hand van een voorbeeld, namelijk het koppelen van twee bestanden.

9.2 Voorbeeld: koppelen van twee bestanden

Indien aan verschillende statistici wordt gevraagd wat zij onder *koppelen* verstaan, dan is de kans groot dat zij verschillende antwoorden geven. Het zal blijken dat ditzelfde begrip door sommigen met een standaard processtap wordt geassocieerd en door anderen met een standaard proces. In deze subparagraaf zullen beide opvattingen worden gemodelleerd.

9.2.1 Achtergrond en definitie van koppelen

Het doel van een koppelproces is *het verkrijgen van een relevante en betrouwbare (micro)dataset met informatie over de betrouwbaarheid*. Iets preciezer gezegd, *koppelen* draagt bij aan de relevantie door variabelen bij elkaar te brengen, en draagt bij aan de betrouwbaarheid en informatie daarover door dekkingsfouten te meten en daarvoor te corrigeren. Uitgaande van twee datasets, waarvan de één recipiënt is en de ander donor, wordt dit gedaan door

- de eenheden in de donor dataset te identificeren met eenheden in de recipiënt dataset op basis van een gemeenschappelijke set van variabelen en een matchingscriterium bij deze set,
- het aantal positieve identificaties te meten (eenheid is in zowel in recipiënt als in donor dataset),
- het aantal gemiste identificaties te meten (eenheid zit wel in recipiënt maar niet in donor dataset),
- het aantal ontbrekende identificaties te meten (eenheid zit niet in recipiënt maar wel in donor dataset) en eventueel toe te voegen aan de recipiënt dataset,
- voor de positieve (en de toegevoegde ontbrekende) identificaties een selectie van de variabelen uit de donor dataset toe te voegen aan een selectie van de variabelen in de recipiënt dataset.

Deze oplossing gaat uit van de ontwerpvoorwaarden dat 1) de populatieafbakeningen van recipiënt en donor dataset aan elkaar gelijk zijn en 2) beide datasets een gemeenschappelijke set van variabelen hebben.

Merk op het aantal gemiste identificaties een indicatie is voor overdekking in de recipiënt dataset, en het aantal ontbrekende identificaties een indicatie voor onderdekking in de donor dataset.

De gemeenschappelijke variabelen in de sets waarop geïdentificeerd wordt, noemen we koppelvariabelen. Idealiter zijn de koppelvariabelen voor beide datasets identiek gedefinieerd en geven zij perfecte sleutels op basis waarvan een identificatie kan plaatsvinden. In de praktijk zijn koppelvariabelen vaak wel vergelijkbaar maar niet identiek gedefinieerd. Bovendien kunnen er ontbrekende en/of foute waarden zijn en kleine tijdsverschillen, waardoor de uiteindelijke sleutels imperfect zijn. Op basis van een matchingscriterium zal dan een identificatie moeten plaatsvinden.

Stel dat twee datasets over een populatie van personen op basis van vijf variabelen worden gekoppeld. Voor het recipiënt bestand zijn dit de variabelen BSN-nummer, naam, woonplaats, geslacht en leeftijd. Voor het donorbestand zijn dit de variabelen BSN-nummer, naam, woonplaats, geslacht en geboortedatum. Stel verder dat de variabelen BSN-nummer en geslacht in beide bestanden uniek gedefinieerd zijn (mogen wel missings en fouten bevatten) en dat de variabelen naam en woonplaats (open) tekstvariabelen zijn (mogen ook missings en fouten bevatten).

Mits de BSN-nummers in beide datasets nonmissing zijn en geen fouten bevatten, wordt bij voorkeur op BSN-nummer gekoppeld. Daartoe wordt de variabele in beide datasets gecontroleerd (met de modulo 11-regel). Aan de hand van de geldige BSN-nummers kan vervolgens een daadwerkelijke koppeling plaatsvinden met als matchingscriterium dat de nummers van donor en recipiënt identiek moeten zijn.

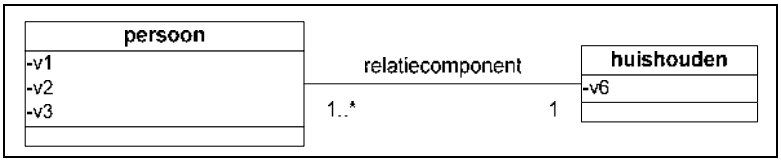
Indien het koppelrendement op de variabele BSN-nummer vanwege missing en/of foute BSN-nummers onvoldoende is, kan een tweede koppelingsslag plaatsvinden, maar nu aan de hand van de variabelen naam, woonplaats, geslacht en leeftijd/geboortedatum. Om de (open) tekstvariabelen naam en woonplaats bij de koppeling te kunnen gebruiken worden beide variabelen aan de hand van een kennistabel naar dezelfde (gesloten) categorieën herleid. Dit kan worden opgevat als het typeren (van variabelen). Om op de variabelen leeftijd/geboortedatum te kunnen koppelen wordt in het donorbestand de variabele leeftijd uit geboortedatum afgeleid. Eventueel is het nodig om de variabelen geslacht te hercoderen naar dezelfde (standaard)codering. Mits de getypeerde, afgeleide en gehercodeerde variabelen geen missings en/of fouten bevatten – daartoe worden zij gecontroleerd – kunnen zij bij de koppeling worden gebruikt. Als criterium kan worden gehanteerd dat de (getypeerde) waarden van naam en woonplaats identiek moeten zijn en dat daarnaast nog minstens een match moet worden gevonden bij geslacht of leeftijd.

Er zijn drie mogelijkheden: er wordt geen bijpassende donor gevonden, er wordt precies één bijpassende donor gevonden, of er worden meerdere bijpassende donoren gevonden. Bij de eerste mogelijkheid spreken we van een gemiste identificatie, bij de tweede en derde mogelijkheid van een positieve identificatie. De kans dat er meerdere bijpassende donoren worden gevonden en dat bij een positieve identificatie de verkeerde donor is toegewezen kan (in de ontwerpfase) worden ingeschat aan de hand van een frequentieverdeling waarbij de klassen gevormd zijn door de set van koppelvariabelen.

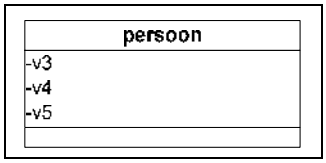
Om te kunnen koppelen zijn dus als input twee datasets nodig (recipiënt en donor) en een matchingscriterium. Beide datasets mogen een relatief ingewikkeld productontwerp hebben waarbij meerdere objecttypen en relaties daartussen kunnen voorkomen.

Voorwaarde is dat beide productontwerpen een gemeenschappelijk objecttype hebben met dezelfde populatieafbakening en een gemeenschappelijke set van variabelen, zie figuur C-13a en C-13b.

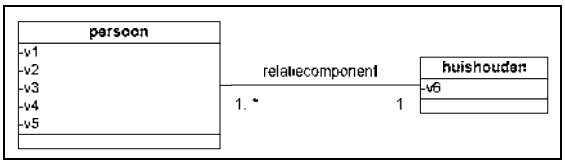
Figuur C-13a: Voorbeeld logisch datamodel van de recipiëntset



Figuur C-13b: Voorbeeld logisch datamodel van de donorset



Figuur C-13c: Voorbeeld logisch datamodel van het koppelresultaat



De output van het koppelen betreft een dataset met een populatieafbakening die gelijk is aan die van de recipiënt en donor dataset, en met een selectie van variabelen uit zowel de recipiënt als de donor dataset (zie figuur C-13c). Merk op dat de instantie van de afbakening gelijk genomen kan worden aan die van de recipiënt (left join), aan de vereniging tussen recipiënt en donor (outer join) of aan de doorsnede tussen recipiënt en donor (inner join). Deze keuze is een ontwerpbeslissing en is vastgelegd in het productvoorschrift. Het is een afweging tussen kwaliteitseisen met betrekking tot de dekking en kwaliteitseisen met betrekking tot missing data.

9.2.2 Typische handelingen bij koppelen

Stel we willen twee datasets koppelen op basis van een set van (bijna) gemeenschappelijke variabelen. Uitgaande van de populatieafbakening van de recipiënt dataset zijn typische handelingen daarbij *datasets inlezen,* ***recipiënt dataset en donor dataset selecteren****, variabelen (her)coderen, variabelen en/of typeren,* ***de set van koppelvariabelen selecteren****,* ***de koppelvariabelen controleren op waardebereik****,* ***donoren identificeren****,* ***de set van te koppelen variabelen selecteren voor de output*** en *wegschrijven van de output*.

De bovengenoemde handelingen die vet gedrukt zijn betreffen typische koppelhandelingen. Deze handelingen laten we in ieder geval deel uitmaken van de standaard processtap *koppelen*. Ten behoeve van de monitoring van de kwaliteit van de

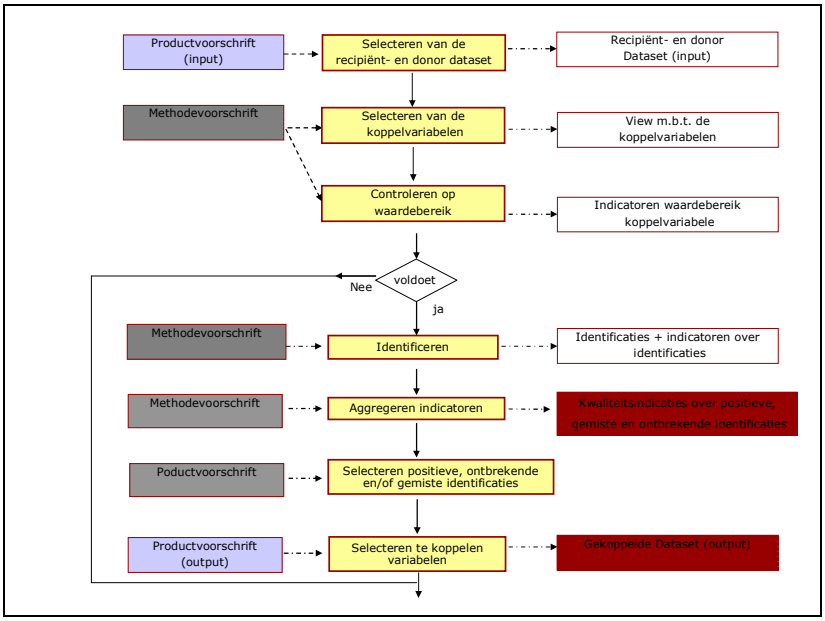
koppeling kiezen we er tevens voor een extra handeling op te nemen, namelijk het *aggregeren van de indicatoren* die het soort identificatie aangeven (positief, gemist of ontbrekend).

9.2.3 Koppelen als standaard processtap

In figuur C-14 is de algemene structuur van de standaard processtap *koppelen* weergegeven. Het ontwerp van deze processtap bestaat uit de volgende onderdelen:

1. ontwerp van de productvoorschriften
    a. het ontwerp van het outputbestand (populatieafbakeningen, sets van variabelen en kwaliteitseisen met betrekking tot over- en onderdekking), zie ook figuur C-13c.
    b. uit de kwaliteitseisen van (a) volgt het ontwerp van de instantie van de populatieafbakening (met of zonder de gemiste en/of ontbrekende identificaties).
    c. het ontwerp van het recipiënt- en donor inputbestand (populatieafbakeningen, sets van variabelen en kwaliteitseisen met betrekking tot over- en onderdekking), zie ook figuren 17a en 17b.
2. het ontwerp van de methodevoorschriften
    a. het ontwerp van de koppelvariabelen (en hun waardebereiken) en het matchingscriterium,
    b. het ontwerp van kwaliteitsindicatoren – deze zullen mede gebaseerd zijn op de aantallen gemiste en ontbrekende identificaties – voor het meten van onder- en overdekking,
    c. de norm voor de kwaliteitsindicatoren in relatie tot de gestelde kwaliteitseisen met betrekking tot over- en onderdekking,

Figuur C-14: typische structuur van de standaard processtap *koppelen*



Aan een matchingscriterium ligt vaak impliciet dan wel expliciet een statistische methode ten grondslag. Voor de volledigheid zal deze methode in subparagraaf 9.2.4 hieronder ter sprake komen.

Indien we de standaard processtap tijdelijk als een black box opvatten dan zijn de product- en de methodevoorschriften de input van de processtap die deze stap inhoudelijk sturen. Indien we in de black box kijken dan zien we dat de uitvoering van de standaard processtap *koppelen* uit zes soorten handelingen bestaat:

1. het selecteren van de recipiënt- en donor dataset op basis van het productvoorschrift (van de input),
2. het selecteren van de koppelvariabelen op basis van het methodevoorschrift,
3. het controleren op waardebereik op basis van het methodevoorschrift,
4. het identificeren van donoren op basis van een matchingscriterium in het methodevoorschrift,
5. het aggregeren van de matchingsindicatoren op basis van het methodevoorschrift,
6. het selecteren van de positieve, ontbrekende en/of gemiste identificaties en de te koppelen variabelen op basis van het productvoorschrift (van de output).

Om de verschillende handelingen te kunnen uitvoeren zijn niet alleen de inhoudelijk sturingsvoorschriften nodig om de handelingen te sturen, maar zijn ook het recipiënt- en donor bestand als input nodig. De output van de processtap is een gekoppelde dataset plus kwaliteitsinformatie ten aanzien van dekking ten gevolge van de koppeling.

Vaak wordt onder koppelen alleen het hierboven genoemde geheel van soorten handelingen verstaan.

9.2.4 Statistische methode[10]

In paragraaf 3.1 zijn bij het ingekaderde voorbeeld twee matchingscriteria beschreven. Het eerste matchingscriterium betrof een criterium om BSN-nummers met elkaar te vergelijken. Volgens dit criterium vindt er een positieve identificatie plaats als BSN-nummer van recipiënt en donor identiek zijn. Het tweede matchingscriterium betrof een criterium om de getypeerde, afgeleide en gehercodeerde variabelen Naam, Woonplaats, Geslacht, Leeftijd onderling te vergelijken. Volgens dit tweede criterium vindt er alleen een positieve identificatie plaats als de (getypeerde) waarden van Naam en Woonplaats identiek zijn en bovendien nog minstens een match moet worden gevonden bij Geslacht of Leeftijd.

Gegeven de set van koppelvariabelen $\{v_1,...,v_K\}$, kunnen beide criteria worden opgevat als een instelling van een algemene matchingsregel:

[10] Voor het onderwerp koppelen zijn nog geen methoden in de Methodenreeks van DMK opgenomen.

- Definieer per koppelvariabele v_k een verschilmaat D_k tussen de (eventueel getypeerde, afgeleide, en/of gehercodeerde) waarde van de recipiënt en de (eventueel getypeerde, afgeleide, en/of gehercodeerde) waarde van de donor.
    - Voor nominale variabelen is D_k bijvoorbeeld gelijk aan 0 als de waarden identiek zijn en anders 1.
    - Voor intervalvariabelen is D_k bijvoorbeeld gelijk aan 0 als de waarden minder dan x% van elkaar afwijken en anders 1.
- Definieer nu voor de gehele set $\{v_1,...,v_K\}$ een gewogen verschilmaat: $D = \sum_k w_k D_k$, waarbij het gewicht w_k het belang van iedere koppelvariabele aangeeft: $\sum_k w_k = 1$.
- Besluit tot een positieve identificatie als $D \leq k$.

Het is duidelijk dat het eerste criterium – BSN-nummer van donor en recipiënt moeten identiek zijn – hier een parametrisatie van is, door het gewicht voor de variabele BSN-nummer gelijk aan 1 te nemen en $k = 0$ te stellen.

Met iets meer creativiteit kan ook het tweede criterium – Naam en Woonplaats moeten identiek zijn en minstens Geslacht of Leeftijd – als een instelling (parametrisatie) van de matchingsregel worden verkregen. Neem bijvoorbeeld het gewicht voor de variabelen Naam en Woonplaats elk gelijk aan $5/12$ en het gewicht voor de variabelen Geslacht en Leeftijd elk gelijk aan $1/12$. Voor de keuze $k = 1/12$ wordt dan het tweede criterium verkregen.

Zoals gezegd dient in de ontwerpfase door de statisticus het matchingscriterium te worden aangegeven in het methodevoorschrift. Indien dit methodevoorschrift toelaat om het matchingscriterium via parameters als w_k en k aan te geven, dan noemen we het methodevoorschrift instelbaar (of parametriseerbaar). De statisticus hoeft niet iedere keer een volledig nieuw methodevoorschrift te ontwerpen, maar (her)gebruikt de statistische methode die ten grondslag ligt aan de parametrisering voor het matchingscriterium.

9.3 Standaard proces

Als ontwerpvoorwaarde voor de standaard processtap *koppelen* dienen de koppelvariabelen in recipiëntset en donorset identiek gedefinieerd te zijn, i.e. dezelfde variabeledefinities en dezelfde waardebereiken, te hebben. Een willekeurige recipiënt- en donorset zullen niet aan deze voorwaarden voldoen en beide datasets zullen daartoe ‘voorbereid’ moeten worden.

- Op basis van de oude set van koppelvariabelen zal een nieuwe set moeten worden gehercodeerd, afgeleid en/of getypeerd.

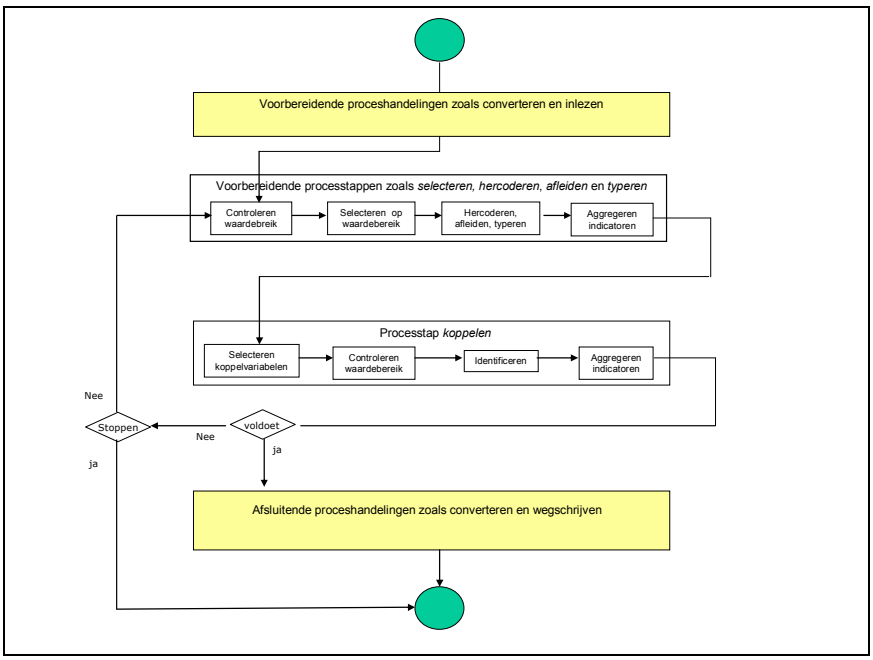
De voorbereidende handelingen *(her)coderen, afleiden en typeren van variabelen* zijn dan ook geen typische koppelhandelingen, maar noodzakelijke handelingen om aan de ontwerpvoorwaarden van de standaard processtap *koppelen* te kunnen voldoen.

Ook deze handelingen zullen worden aangeduid als standaard processtappen, omdat zij alle drie, evenals koppelen, geassocieerd kunnen worden met het (generieke) doel om een relevante en betrouwbare (micro)dataset te verkrijgen. Bij *hercoderen* wordt dit doel verkregen door variabelen te hercoderen, bij *afleiden* door variabelen op een deterministische manier af te leiden uit andere variabelen, en bij typeren door variabelen met open antwoordcategorieën te herleiden tot variabelen met gesloten categorieën aan de hand van een zogenaamde kennistabel.

Om de standaard processtappen te kunnen starten zijn bovendien nog handelingen nodig als *het inlezen van datasets* en/of *het transformeren van het formaat van een dataset*. Om de standaard processtappen te kunnen afsluiten dienen de verkregen outputs te worden weggeschreven en kan het mogelijk zijn dat hiervoor weer een formaattransformatie nodig is.

In figuur C-15 is een typische structuur van dit standaard proces *koppelen* schematisch weergegeven. Het proces is opgebouwd uit een aantal standaard processtappen, namelijk *afleiden van variabelen, (her)coderen, typeren* en *koppelen,* waarvan de standaard processtap *koppelen* de cruciale stap is van het proces. De andere stappen zijn voorbereidend. Voor, na en tussen de standaard processtappen zitten handelingen die niet noodzakelijk zijn voor het doel, maar wel voor de uitvoer van het proces.

Figuur C-15: typische structuur van de het standaard proces koppelen



Er is een iteratietrigger ingebouwd die op basis van een norm de processtap *koppelen* met haar voorbereidende processtappen laat itereren, waarbij bij iedere iteratie een andere specificatie van het methodevoorschrift genomen kan worden. Er is eveneens een trigger

ingebouwd om het aantal iteraties te beperken en het proces te stoppen. Deze trigger kan zijn op basis van een maximaal aantal iteraties, maar ook op basis van een tijdsplanning.

Merk op dat onder koppelen vaak niet alleen de processtap wordt verstaan, maar dit gehele proces.

Formeel definiëren we een standaard proces als een proces met een expliciet doel dat wordt gerealiseerd met een expliciet geformuleerde (generieke) methodologische oplossing. Het bestaat uit één of meer standaard processtappen en/of handelingen, en onderscheidt zich van een standaard processtap vanwege voor- en nabereidende handelingen en eventuele iteraties. Een standaard proces wordt gestuurd door procesvoorschriften ten aanzien van zowel de volgorde van de processtappen en/of handelingen als de planning, zie paragraaf 9.4.

9.4 Sturingsvoorschriften: procesmetadata

9.4.1 Inhoudelijke voorschriften voor standaard processtappen

Bij de uitvoering van de standaard processtap dient op verschillende plekken de processtap inhoudelijk gestuurd te worden. Voor een standaard processtap in het statistische verwerkingsproces onderscheiden we twee (typen) van inhoudelijke sturingsvoorschriften.

- productvoorschrift/productontwerp: dit voorschrift komt overeen met de productbeschrijving. Het beschrijft de in- en outputproducten van de processtap in termen van conceptuele metadata, namelijk door middel van een beschrijving van een populatie en een opsomming van variabelen. Ook stelt een dergelijk voorschrift kwaliteitseisen aan input- en outputproducten
- methodevoorschrift: dit voorschrift komt overeen met de gebruiksaanwijzing. Het beschrijft de methodologische oplossing die moet worden toegepast in de processtap om
    - het outputproduct te maken
    - de kwaliteit van het outputproduct te meten (kwaliteitsindicatoren)
    - de kwaliteitsindicatoren te relateren aan de kwaliteitseisen (normen)

Deze inhoudelijke sturingsvoorschriften komen van buiten het verwerkingsproces en zijn tot stand gekomen in het ontwerpproces.

9.4.2 Procesmatige voorschriften voor standaard processen

De uitvoering van het gehele standaard proces dient procesmatig gestuurd te worden. We onderscheiden twee (typen) van procesmatige voorschriften:

- procesvoorschrift/procesontwerp: dit voorschrift geeft aan

- welke standaard processtappen en/of handelingen in welke volgorde met welke (versies) van product- en methodevoorschriften in het standaard proces moet worden uitgevoerd
- welke procesindicatoren nodig zijn om de kwaliteit (effectiviteit en snelheid) van het proces te meten
- welke normen bij de procesindicatoren nodig zijn ten behoeve van het beëindigen, (bij)sturen en/of verbeteren van het proces

- planningsvoorschrift: dit voorschrift geeft aan wanneer, met welke regelmaat en met welke resources een standaard proces moet worden uitgevoerd.

Een procesvoorschrift/procesontwerp bestaat daarmee uit een geordende verzameling van (geversioneerde) productvoorschriften, methodevoorschriften en procesindicatoren met bijbehorende normen.

Bij het plannen van het standaard proces dienen de start- en einddata voldoende uit elkaar te liggen zodat de op basis van het procesontwerp geschatte doorlooptijden, inclusief marges, erin passen.

Ook de procesmatige voorschriften komen van buiten het verwerkingsproces. Zij zijn eveneens tot stand gekomen in het ontwerpproces. Merk op dat het procesontwerp kan worden vastgelegd in tools als MAVIM.

9.4.3 Procesmetadata en voorschrijvende conceptuele metadata

Methodevoorschriften vormen samen met de procesvoorschriften de procesmetadata. De methodevoorschriften vormen het inhoudelijke deel van de procesmetadata en de procesvoorschriften het procesdeel. Door dit onderscheid te maken wordt het inhoudelijke deel, de ‘know’, van het procesmatige deel, de ‘flow’, gescheiden. Productvoorschiften vormen de (voorschrijvende) conceptuele metadata.

9.5 Beheer van sturingsvoorschriften en change management

Figuur C-16 geeft nogmaals het domein *verwerken* op hoofdlijnen weer, waarbij nu de ‘losse’ handelingen zijn vervangen door een standaard proces en de ‘bak’ met productontwerpen bestaat uit de sturingsvoorschriften waarop structuur is aangebracht. De sturingsvoorschriften zullen tijdens de implementatie van het proces (de bouw) uitmonden in fysieke regelsets. Deze fysieke regelsets hebben uiteraard dezelfde structuur als sturingsvoorschriften.

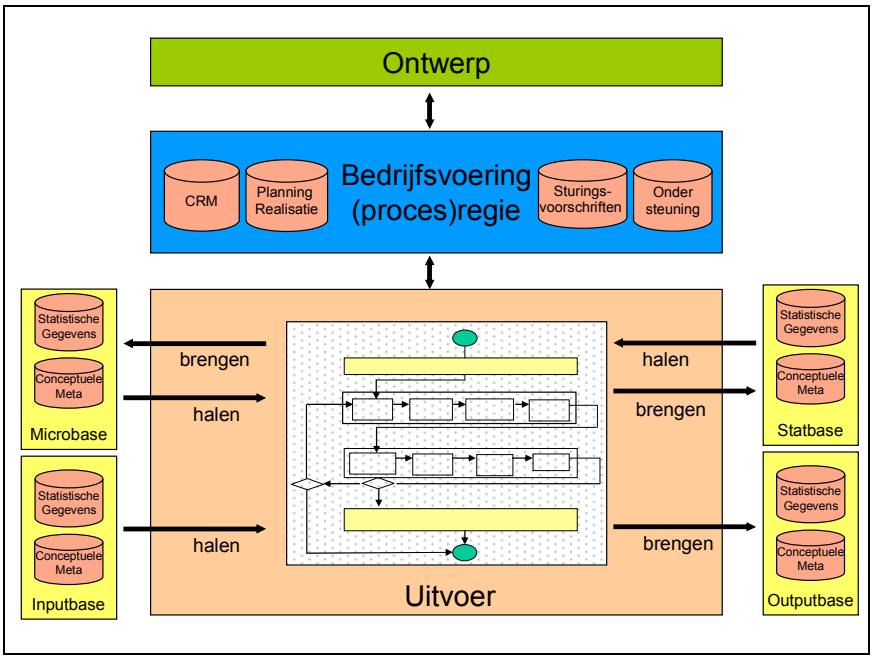
Tijdens het ‘draaien’ zullen, conform het proces- en planningsvoorschrift, op het juiste moment de juiste regelsets voor de productvoorschriften, de juiste regelsets voor de methodevoorschriften en de juiste regelsets voor de kwaliteitsindicatoren toegepast moeten worden. Dit is in deze cursus de interpretatie ‘regelsturing’. Voor grote verzamelingen regelsets, die in de loop der tijd kunnen veranderen, is een goed beheer- en releasebeleid van belang om iedere keer ‘de juiste’ regelset op het juiste moment te

‘draaien’. Zonder het wiel opnieuw te willen uitvinden behelst een change management procedure de volgende stappen

- Verzoek tot verandering t.a.v. één of meer sturingsvoorschriften aan een change management board (deze bestaat minstens uit een eigenaar van de voorschriften, één of meer gebruikers daarvan en één of meer afnemers van de te realiseren rustpunten).
- Na het prioriteren van de verzoeken volgt het ontwikkelen en uitproberen van de fysieke regelsets (in een ontwikkelomgeving).
- Vervolgens wordt de ontwikkelde fysieke regelset getest in een acceptatieomgeving, op basis waarvan de change management board haar fiat kan geven (het testen en fiatteren is inclusief het ontwerpen van testcases).
- Na fiattering wordt de nieuwe regelset in productie genomen (en geversioneerd).

Voorafgaand aan de change management procedure zijn er prikkels/aanleidingen om tot een change management verzoek te komen. Een dergelijke aanleiding kan bijvoorbeeld komen uit de reguliere productie, waarbij tijdens het monitoren blijkt dat het proces stagneert, dat er (te) veel uitval is, dat de inputs anders zijn dan verwacht, dat de afgesproken kwaliteit keer op keer niet gehaald wordt, etc. In een apart proces dienen de productieproblemen nader onderzocht te worden – ook een dergelijk onderzoeksproces dient te worden ingericht – en dat kan uiteindelijk leiden tot een verzoek voor verandering.

Figuur C-16: Nogmaals het verwerkingsdomein in hoofdlijnen



9.6 Inventarisatie van standaard processtappen

Momenteel is er nog geen bibliotheek met typische standaard processtappen waaruit (standaard) processen kunnen worden opgebouwd. Wel zijn er typische processtappen besproken die kandidaat voor de bibliotheek zijn. Deze typische processtappen zijn

- steekproeftrekken,
- koppelen,
- typeren,
- afleiden (van variabelen),

waarbij de eerste kandidaat processtap specifiek voor het domein *waarnemen* lijkt te zijn, dat wil zeggen, vooral in het domein *waarnemen* worden toegepast. Dat is overigens geen ijzeren wet. De idee achter een standaard processtap is dat het autonome bouwstenen (componenten) zijn waarmee processen kunnen worden opgebouwd. Ze vertegenwoordigen een statistische functionaliteit die in een standaard proces ingezet kan worden als oplossing voor een statistisch doel.

Bovenstaande kandidaatlijst kan nog uitgebreid worden met andere voor de hand liggende kandidaten. Ter illustratie noemen we

- afleiden (van eenheden en variabelen),
- gaafmaken (inclusief selectief detecteren van fouten en imputeren),
- schatten (van populatieparameters via weging met regressie),
- schatten (van populatieparameters via kleine-domein-methoden),
- detecteren van uitbijters,
- schatten (van varianties met delta-methoden),
- inpassen (van tabellen via Lagrange),
- inpassen (van tijdreeksen via …),
- corrigeren (van trendbreuken),
- statistisch beveiligen (van tabellen),
- statistisch beveiligen (van microdata).

De idee is om de verschillende kandidaten als standaard processtap uit te werken en op te nemen in de bibliotheek.

De hierboven opgesomde lijst is niet compleet. Dat kan ook niet omdat er voortdurend nieuw generieke oplossingen worden bedacht waaruit nieuwe standaard processtappen voortkomen. Deze kunnen worden toegevoegd aan de oude of de oude vervangen. Kortom, de lijst is levend.

9.7 Relatie met standaard toolset

Idealiter wordt een verwerkingsproces (op papier en/of in MAVIM) zodanig ontworpen dat de in de bibliotheek gedefinieerde processtappen herkenbaar en gespecificeerd zijn. Bij daadwerkelijke implementatie kunnen hier vervolgens wel of geen standaard tools voor beschikbaar zijn. Indien er wel een standaard tool voor beschikbaar is dan kan bovendien altijd nog de keuze worden gemaakt, mits onderbouwd, hier wel of geen gebruik van te maken.

Bij de keuze van een tool spelen onder meer de volgende twee criteria een rol.

- Wat is de instelbaarheid van het methodevoorschrift, bijvoorbeeld ten aanzien van het matchingscriterium?
- Wat is de scope van het te implementeren proces? Is de scope beperkt tot een handeling, tot een standaard processtap, of betreft de scope een heel standaard proces, inclusief eventuele iteraties.

Aan de hand van bovengenoemde criteria onderscheiden we twee vormen van generiekheid van een standaard tool:

- Een standaard tool kan *generiek* zijn wat betreft de instelbaarheid van het methodevoorschrift, maar *specifiek* (beperkt) ten aanzien van de scope. De tool bestrijkt bijvoorbeeld alleen een standaard processtap.
- Een standaard tool kan specifiek ten aanzien van de instelbaarheid van het methodevoorschrift, maar *generiek* (breed) ten aanzien van de scope. De tool bestrijkt het gehele standaard proces, maar beperkt zich bij bijvoorbeeld *koppelen* tot het koppelen van datasets op grond van unieke sleutels.

Indien het methodevoorschrift bij bijvoorbeeld *koppelen* niet instelbaar hoeft te zijn en bovendien het matchingscriterium eenvoudig is, dan zijn er meerdere standaard tools waarmee de standaard processtap *koppelen* relatief eenvoudig geïmplementeerd kan worden. Denk daarbij bijvoorbeeld aan de generieke standaard tools SPSS en Manipula. Eventueel kunnen de voorbereidende standaard processtappen *afleiden van variabelen* en *hercoderen* ook met deze tools worden geïmplementeerd en zelfs het gehele standaard proces.

Problematischer wordt het indien het methodevoorschrift wel instelbaar moet zijn en de koppelvariabelen eerst getypeerd moeten worden. Tools als Trillium komen dan wellicht meer in aanmerking. Echter, het is goed mogelijk dat dit soort tools niet het gehele standaard proces afdekken, waardoor er aanvullende tools nodig zijn en er vervolgens ‘integratie’ tussen de tools nodig is.

9.8 Voordelen van standaardisatie

We zijn de paragraaf begonnen met het stellen van een aantal voorwaarden aan een ontwerp van een proces vanuit de Business Architectuur. Processen moeten zijn effectief en tijdig, efficiënt en flexibel, en transparant en reproduceerbaar.

Via de concepten *standaard processtappen* (met productvoorschriften en methode-voorschriften) en *standaard processen* (met procesvoorschriften en planningsvoorschriften) hebben we getracht structuur in de processen aan te brengen, waarbij doel, oplossing en output van elkaar gescheiden zijn.

Het aanbrengen van deze structuur is niet een doel op zich, maar een middel, zeg oplossing, om het ontwerp van de processen te verbeteren.

- de processen worden effectiever doordat doel en output gescheiden zijn, het doel via de productvoorschriften expliciet is ingebracht en de output op dit doel wordt gemonitord
- De processen worden efficiënter (goedkoper) doordat de generieke componenten herbruikbaar zijn, en deze componenten daar waar mogelijk gebruik maken van standaard statistische methoden en standaard IT-oplossingen.
- De processen worden flexibeler doordat een specificatie van een generieke component kan worden aangepast zonder 'last' te hebben van andere componenten. Vanwege de autonomie kunnen componenten in hun geheel worden vervangen.
- De processen worden transparanter doordat ze zijn opgebouwd aan de hand van een generieke structuur met herkenbare generieke componenten.
    - Daar waar mogelijk liggen herkenbare statistische methoden ten grondslag aan de generieke componenten (standaard methodologie oplossing).
    - Daar waar mogelijk zijn de generieke componenten geïmplementeerd met standaard tooling (standaard IT-oplossing).
- De reproduceerbaarheid van de processen wordt vergroot via een release en change management beleid op de sturingsvoorschriften.

Naast bovengenoemde voordelen kan worden genoemd dat processen makkelijker documenteerbaar worden, doordat kan worden gerefereerd naar de documentatie van de generieke componenten.

10. Het domein statistische informatieontwikkeling en publicatie

Het doel van statistische informatieontwikkeling en publicatie is om de statistische gegevens als eindproduct te 'upgraden' tot statistische informatie en deze informatie voor externe gebruikers in een presentatievorm toegankelijk te maken.

Deze in een presentatievorm gegoten en toegankelijk gemaakte statistische informatie noemen we het eindproduct van het CBS. Figuur C-17 geeft op hoofdlijnen het domein *statistische informatieontwikkeling en publicatie* weer.

De Business Architectuur van het CBS stelt dat het domein *verwerken en statistische beveiliging* ontkoppeld is van het domein *informatieontwikkeling en publicatie* met de outputbase als ontkoppelpunt. Het outputdomein start dan ook bij de statistische gegevens die door het verwerkingsdomein via de outputbase beschikbaar zijn gesteld. Op basis van

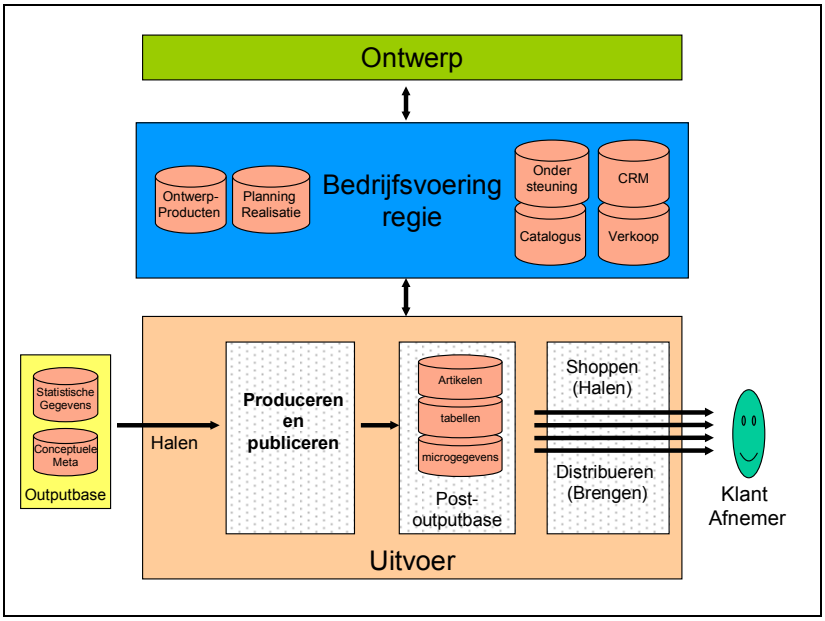
deze statistische gegevens worden door het outputdomein drie soorten eindproducten geproduceerd en gepubliceerd, namelijk artikelen, tabellen en microdata.

Artikelen worden als eindproduct verkregen door uit de outputbase de gewenste statistische gegevens te selecteren, hierop eventuele statistische analyses te plegen en over het resultaat een artikel te schrijven. Voor artikelen gelden vaste presentatievormen (layouts).

Op soortgelijke wijze worden tabellen als eindproduct verkregen. Uit de outputbase worden de gewenste statistische gegevens geselecteerd, hierbij worden vaak toelichtingen geschreven en het resultaat wordt in een tabel samengesteld. Voor tabellen gelden vaste presentatievormen (layouts).

Ten slotte worden de gewenste microgegevens uit de outputbase geselecteerd en zonodig getransformeerd naar het gewenste formaat. Deze microgegevens als eindproduct zijn bijna altijd maatwerk voor niet-anonieme klanten.

Figuur C-17: Het outputdomein in hoofdlijnen



De post-outputbase in figuur C-17 betreft een virtuele verzamelplaats (CBS-winkel) van waaruit de eindproducten voor externe klanten toegankelijk zijn gemaakt. Hiertoe zijn meerdere distributiekanalen op de post-outputbase aangesloten, zoals het internet, e-mail, on-site, etc. Bij het distribueren neemt het CBS het initiatief en bij het virtueel winkelen de klant.

De Business Architectuur van het CBS stelt dat *produceren en publiceren* ontkoppeld is van *shoppen en distribueren* met de post-outputbase als (virtueel) ontkoppelpunt.

10.1 Produceren en publiceren

De Business Architectuur van het CBS stelt tevens dat alle statistische informatie in de (virtuele) post-outputbase terug te voeren moet zijn op statistische gegevens in de

outputbase. De scope van het produceren en publiceren is daarmee beperkt tot het duiden, analyseren, interpreteren, presenteren en communiceren van statistische gegevens. De statistische gegevens zelf blijven onveranderd. Daartoe stelt de Business Architectuur de voorwaarden

- Eens in de outputbase, altijd in de outputbase. Met andere woorden, indien statistische gegevens als rustpunt intern zijn vrijgegeven voor informatieontwikkeling, dan kunnen deze gegevens niet zomaar teruggetrokken worden.
- Statistische gegevens in de outputbase voldoen aan de kwaliteitseisen. Zij zijn dus betrouwbaar en samenhangend, en het verwerkingsdomein heeft hiertoe voor gezorgd.

Mocht overigens in het outputdomein blijken dat de gegevens toch niet aan de kwaliteitseisen voldoen, dan wordt dit teruggekoppeld naar de verantwoordelijke eigenaar in het verwerkingsdomein.

Uit eenzelfde set statistische gegevens kunnen meerdere eindproducten geproduceerd worden, met bijvoorbeeld verschillende duidingen in verschillende talen voor verschillende doelgroepen te bereiken via verschillende distributiekanalen. De Business Architectuur stelt

- de keuze van het soort informatie, inclusief taal en presentatievorm in combinatie met een distributiekanaal, is klantgericht.

Zodra de eindproducten klaar zijn, worden vrijgegeven, zeg gepubliceerd, in de post-outputbase van waaruit ze gedistribueerd worden of van waaruit de klant virtueel kan winkelen.

10.2 Distribueren

Er zijn verschillende kanalen waarlangs het CBS een eindproduct kan distribueren. Een eerste indeling die we hierin willen aanbrengen is het verschil tussen on-line en off-line distributie.

Bij on-line distributie worden de (electronische) eindproducten via electronische kanalen bereikbaar gemaakt. We onderscheiden de volgende vormen

- publieke websites, zoals StatWeb, CBS.nl en specifieke doelgroep gerichte sites.
- on-site afgesloten werkplekken op het CBS.
- E-mail, file transfer, RSS-feeds en SMS.

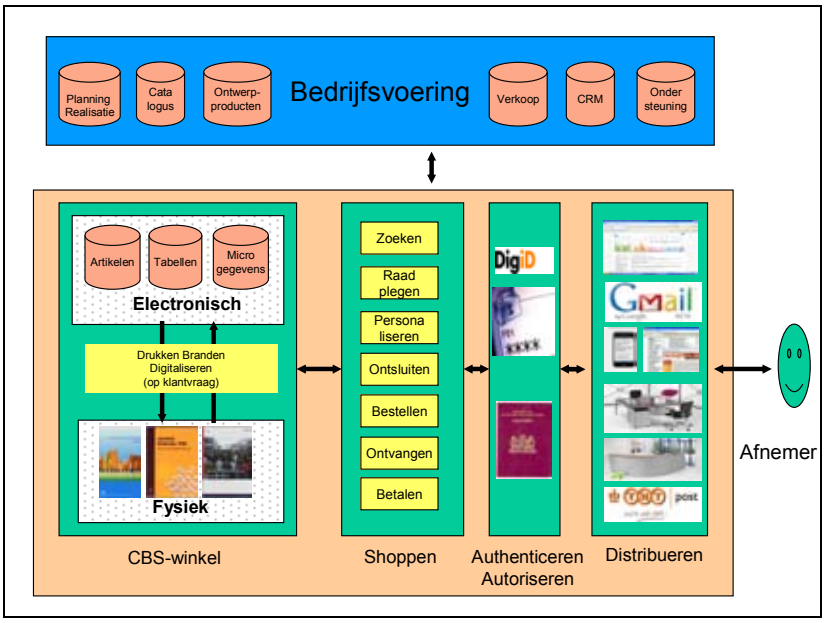
Bij off-line distributie worden de electronische producten eerst fysiek gemaakt door ze bijvoorbeeld in boekvorm te printen, om ze vervolgens per fysieke post te versturen bij de balie te laten afhalen. De Business Architectuur stelt dat de electronische producten leidend zijn.

Naast on-line en off-line distributie kent het CBS ook verbale distributie via de televisie en telefoon.

10.3 Virtueel winkelen

Virtueel winkelen is dat een klant via de publieke websites kunnen rondkijken in de electronische CBS-winkel en producten kunnen kopen, zie figuur C-18.

Figuur C-18: Shoppen en distribueren



Daartoe maken we maken een onderscheid tussen enerzijds anonieme en niet-anonieme klanten en anderzijds tussen statistisch beveiligde en statistisch onbeveiligde informatie.

- Anonieme klanten hebben zich niet bekend gemaakt (geauthentiseerd) en zijn niet geautoriseerd tot bepaalde handelingen. Niet-anonieme klanten hebben zich wel bekend gemaakt en zijn wel geautoriseerd tot bepaalde handelingen.
- Statistisch beveiligde informatie is gebaseerd op statistisch beveiligde gegevens en is toegankelijk voor zowel de anonieme als de niet-anonieme klant. Statistische onbeveiligde informatie bevat onbeveiligde gegevens en is alleen toegankelijk voor daartoe bevoegde klanten.

Zowel anonieme als niet-anonieme klanten kunnen naar producten zoeken, deze producten raadplegen en delen daarvan ontsluiten (downloaden). Anonieme klanten hebben geen toegang tot de gehele winkel, maar alleen tot statistisch beveiligde producten.

Niet-anonieme klanten kunnen tevens maatwerk producten kopen (bestellen, betalen en ontvangen), waarbij ze kunnen kiezen of zij deze producten electronisch willen hebben of fysiek. De Business Architectuur stelt daarbij dat

- de toegankelijkheid van alle statistisch beveiligde eindproducten, dus ook het maatwerk op betaald verzoek, voor alle iedereen gelijktijdig is.

Indien klanten (afnemers) zijn bevoegd, hebben ze tevens toegang tot de voor hen geautoriseerde onbeveiligde producten.

10.4 Magazijn, winkel, etalage en toegangsdeur

We hebben gesproken over de post-outputbase als een virtuele ***CBS-winkel***[11]. Als we in dezelfde beeldspraak blijven dan kunnen we de outputbase als een (intern) ***CBS-magazijn*** opvatten met alle publicabele statistische gegevens van het CBS.

Het verschil tussen het magazijn en de winkel is de volgende. In het magazijn zitten alle ooit door het CBS voortgebrachte publicabele statistische gegevens. Daarmee vervult de outputbase een archieffunctie voor microgegevens en geaggregeerde gegevens. Dagelijks dijt dit magazijn uit doordat er nieuwe statistische gegevens bijkomen. Het domein verwerken brengt haar resultaten als rustpunt naar het magazijn, waarbij een aangeleverd rustpunt meer gegevens kan bevatten dan een specifieke klant nodig heeft, maar niet meer dan alle (potentiële) klanten van dit rustpunt bij elkaar.

Uit één rustpunt kunnen meerdere klant- en/of doelgroepspecifieke eindproducten worden gemaakt die in sterke mate overlappend kunnen zijn. Op één rustpunt gebaseerde eindproducten kunnen bijvoorbeeld verschillen in populatieafbakening, variabelenset, indikkingen en/of coderingen van classificaties, formaat, etc. De eindproducten worden klantvriendelijk toegankelijk gemaakt in de CBS-winkel, waarmee *toegankelijkheid* de belangrijkste functie is van de winkel. De winkel heeft geen archieffunctie en daarom veel minder last van uitdijing, omdat oude producten kunnen worden opgeruimd[12].

Een belangrijk onderdeel van een winkel is een aantrekkelijk ingerichte etalage met een opengeslagen toegangsdeur om de klant te prikkelen om binnen te komen. In de Business Architectuur wordt deze rol vervuld door de publieke websites. Deze sites zijn niet alleen een distributiekanaal, zeg de ***toegangsdeur*** tot de winkel, maar tevens ***etalage***.

---

[11] Het zijn beter gezegd de schappen in de winkel.

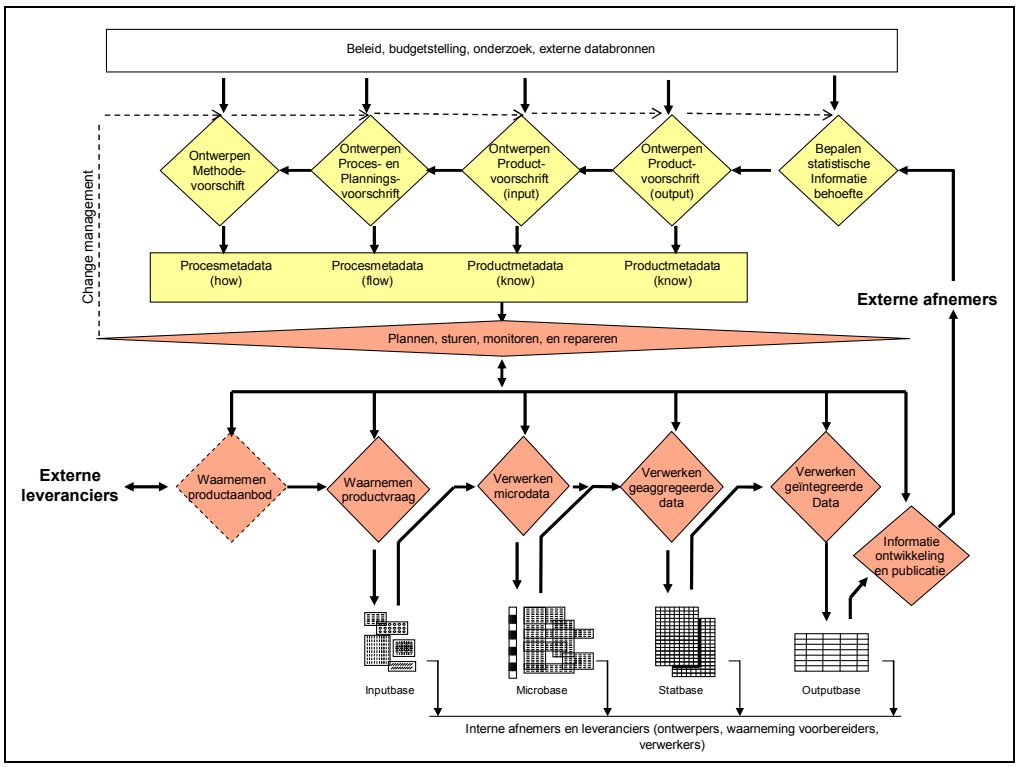
[12] Dit geldt vooral voor de artikelen en tabellen. Microdata komen wellicht helemaal niet in de winkel terecht. Statistische gegevens in tabellen worden gearchiveerd in de outputbase. Artikelen dienen apart gearchiveerd te worden.

Deel D: De Business Architectuur van het CBS

De missie van het Centraal Bureau voor de Statistiek (CBS) is het publiceren van betrouwbare en samenhangende statistische informatie, die inspeelt op de behoefte van de samenleving.

In dit deel vatten we Deel B en C samen in de Business Architectuurplaat van het CBS, welke enigszins is aangepast aan de terminologie van deze cursus, zie figuur D-1. Voor de formele beschrijving van deze plaat verwijzen we naar Huigen (2006a). Tevens gaan we in op de context van het CBS (we kijken naar buiten) en op een aantal generieke diensten (we kijken naar binnen).

Figuur D-1: De Business Architectuurplaat



11. Business architectuurplaat (wat-vraag)

Als je van een afstand naar het CBS zou kunnen kijken, dan zou je een bedrijvigheid zien zoals schematisch weergegeven in figuur D-1. In de figuur kunnen vijf horizontale aspectgebieden worden onderscheiden (beleid, ontwerp, regie, uitvoering en beheer) welke hun toepassinggebied vinden in de drie verticale business domeinen (waarnemen, verwerken en beveiligen, en informatieontwikkeling en publicatie).

Heel in het kort samengevat geldt het volgende. Ontwerp vindt klantgericht plaats, onder beleid en gegeven beschikbare middelen (budgetstelling, methodologische en technologische ontwikkeling, externe bronnen, etc). Het resulteert in voorschrijvende

metadata welke regie en uitvoering stuurt. Regie en uitvoering resulteren in regiedata (proces- en productmetingen en oordelen daarover) respectievelijk statistische data. Regiedata zijn nodig voor de (her)planning, monitoring en (bij)sturing van zowel de lokale processen als de ketens. De statistische data vormen samen met de inhoudelijk definities, interpretaties en presentatievormen de statistische informatie voor de samenleving. Beheer borgt (archiveert) alle resultaten.

11.1 Beleid en beschikbare middelen

Beleid geeft de kaders voor het ontwerp van het statistische proces met betrekking tot zowel de rustpunten als de procesinrichting. Beschikbare middelen (onderzoek, budget, externe databronnen) geven de mogelijkheden. Samen met de klantbehoefte geven beleid en de beschikbare middelen richting aan het ontwerp. De witte rechthoek boven in figuur D-1 representeert het beleid en de beschikbare middelen. De klantbehoefte wordt gerepresenteerd door de gele wieber rechtsboven.

Beleidsresultaten zijn onder meer regels, strategische afspraken (intern, extern), strategische uitspraken op het gebied van bijvoorbeeld waarneming, archivering en statistische beveiliging, speerpunten, jaar- onderzoeksplannen, budgettering, ICT-inzet, etc.

Resultaten van *onderzoek* zijn onder meer wetenschappelijke onderzoeksrapporten waarin statistische methoden zijn beschreven, zie bijvoorbeeld de methodenreeks. Statistische methoden liggen vaak ten grondslag aan de methodevoorschriften bij de standaard processtappen. Resultaten van *externe bronnen* zijn strategische afspraken met registerhouders over het gebruik van registers.

11.2 Ontwerp

Gegeven de statistische informatiebehoefte, het beleid en de beschikbare middelen, resulteert het ontwerp van het statistische proces in de volgende onderdelen (dit zijn de sturingsvoorschriften uit paragaaf 9.4).

- Productvoorschriften voor de outputproducten van dit proces, inclusief kwaliteit (in termen van logische datamodellen, conceptuele metadata en de kwaliteits-metadata).
    - Bij *waarneming* is een onderscheid gemaakt tussen waarneemaanbod en waarneemvraag als outputproduct, waarbij de waarneemvraag overeenkomt met een rustpunt.
    - Bij *verwerking en statistische beveiliging* komen alle outputproducten overeen met rustpunten.
    - Bij *informatieontwikkeling en publicatie* is een onderscheid gemaakt tussen micro-gegevens, tabel en artikel als outputproduct.
- Productvoorschriften voor de inputproducten, inclusief kwaliteit (in termen van conceptuele metadata en kwaliteitsmetadata). Deze inputproducten zijn veelal selecties uit beschikbaar gestelde rustpunten.

- methodevoorschriften (als onderdeel van de procesmetadata) om het ‘gat’ tussen beschikbare inputproducten en de gewenste outputproducten te dichten.
- methodevoorschriften (in termen van kwaliteitsindicatoren en normen) om de kwaliteit van de producten te monitoren.
- procesvoorschriften om de juiste versies van de product- en methodevoorschriften in de juiste combinatie en volgorde voor te schrijven.
- procesvoorschriften (in termen van procesindicatoren en normen) om de kwaliteit van het proces te monitoren.
- planningsvoorschriften om het moment van uitvoeren en de daarbij benodigde middelen voor te schrijven.

De gele wiebers in figuur D-1 stellen lokale omgevingen voor waarin de ontwerpprocessen worden uitgevoerd. De onderliggende gele rechthoek stelt een omgeving voor waarin de ontwerpresultaten zijn geborgd.

Bovengenoemde voorschriften vormen de ontwerponderdelen van een standaard processtap respectievelijk standaard proces, zie paragaaf 9.1 - 9.3. Idealiter beschikt het CBS over een bibliotheek met typische standaard processtappen waaruit de standaard processen kunnen worden opgebouwd, zie paragraaf 9.6. Als deze bibliotheek er is, zal zij deel gaan uitmaken van de Business Architectuur van het CBS.

Bij de daadwerkelijke implementatie van de sturingsvoorschriften ontstaan initiële versies van regelsets. Mocht uit de procesregie blijken dat er in de ontwerpen structurele aanpassingen nodig zijn, dan kunnen deze aanpassingen conform change management worden doorgevoerd. Aanpassingen in de methodevoorschriften zijn het eenvoudigst, daarna komen aanpassingen in de procesvoorschriften en productvoorschriften voor de input. Pas als het niet anders, worden de productvoorschriften voor de output aangepast.

11.3 Regie en uitvoering

Na het ontwerp komt de uitvoering (en de regie daarover). Bij de uitvoering worden de statistische data waargenomen (voor het CBS toegankelijk gemaakt), verwerkt tot statistische gegevens en statistische informatie, statistisch beveiligd en gepubliceerd (voor de samenleving toegankelijk gemaakt).

'Vroeger' – in het tijdperk van de vele stovepipes – waren uitvoeringsprocessen min of meer compleet ingericht van waarnemen tot en met publicatie. Bovendien was er één proceseigenaar (vaak een taakgroepchef) verantwoordelijk voor dit geheel. Veranderingen in het proces of het product konden eenvoudig op hun impact beoordeeld worden. Afhankelijkheden met andere stovepipes waren minimaal of heel los geïmplementeerd.

Vanwege de (externe) druk van efficiencyverbetering en vermindering van administratieve lasten worden de processen tegenwoordig meer als een netwerk ingericht, zie figuur C-3. Door deze (externe) druk zullen steeds vaker ‘stukken’ proces uit

traditionele stovepipes worden gehaald en worden uitgevoerd door collega statistici. Deze collega statistici zijn niet alleen collega's uit andere stovepipes die hetzelfde stuk proces uitvoeren, maar ook de generieke procesdiensten zoals DV en DSC (zie paragraaf 14) die vanwege schaalvoordelen het stuk proces efficiënter kunnen uitvoeren. De onderlinge afhankelijkheden tussen de stovepipes worden daarmee steeds complexer.

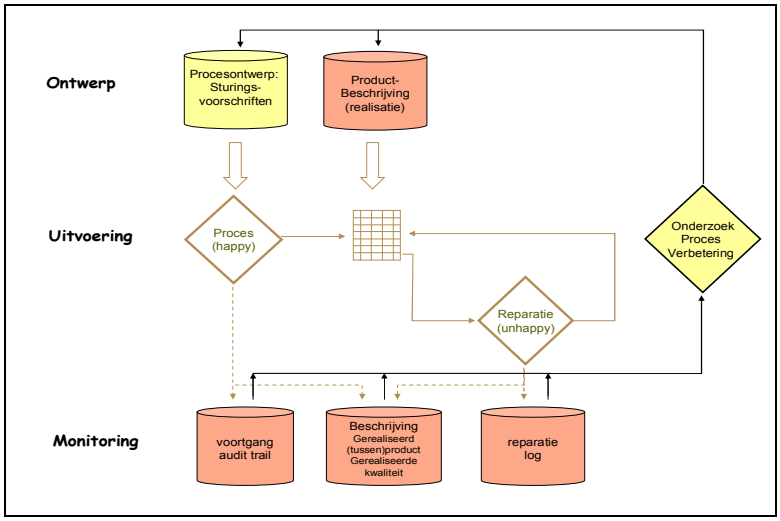
De Business Architectuur van het CBS geeft drie antwoorden op de steeds complexere afhankelijkheden

- introductie van het begrip 'rustpunt'; zie paragraaf 5.2.
- introductie van het begrip 'koppelvlak'; zie paragraaf 5.4.
- onderscheid tussen procesregie en ketenregie; zie ook paragraaf 5.5.

De rode wiebers in figuur D-1 stellen de lokale omgevingen voor waarin de uitvoerings-processen (stukken stovepipe) worden gepland, uitgevoerd en gemonitord. De statistische gegevens die deze processen als rustpunt voortbrengen worden ten behoeve van de uitwisseling ervan 'ontkoppeld' van de lokale omgevingen en geborgd in koppelvlakken. Deze koppelvlakken zijn onderaan in de figuur weergegeven.

De (her)planning, monitoring en (bij)sturing van de uitvoeringsprocessen is niet alleen op het procesniveau, maar ook op ketenniveau. Op procesniveau is de planning meer operationeel, terwijl deze op ketenniveau meer tactisch is.

Figuur D-3: Procesregie



Procesregie is het monitoren van (de happy flow) van een proces (stuk stovepipe) en het zonodig bijsturen van deze flow. De happy flow is conform het procesontwerp en wordt gestuurd vanuit de verzameling sturingsvoorschriften. Het proces heeft een verantwoordelijke eigenaar voor de ontworpen output. Deze output komt overeen met een rustpunt en voldoet dus aan de in paragraaf 5.2 gestelde eisen. Bijsturing van een proces kan een incidentele reparatie inhouden – dit vatten we op als een unhappy flow – of een structurele aanpassing van het ontwerp, zie figuur D-3. Mocht een structurele aanpassing van het ontwerp nodig zijn, dan ontstaat er een nieuwe versie van het procesontwerp en

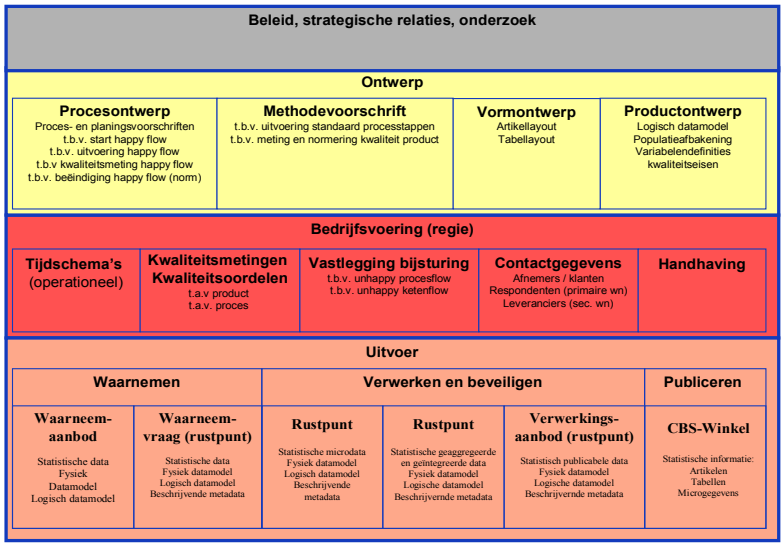
dus een nieuwe happy flow. Het proces resulteert uiteindelijk, eventueel na bijsturing, in een rustpunt waarvan de gerealiseerde inhoudelijke betekenis en de gerealiseerde kwaliteit is beschreven.

Ketenregie is het monitoren van (de happy flow) van een keten en zonodig het nemen van een actie (vaak in de vorm van overleg). De keten is hierbij de verzameling van alle schakels tussen de processen, inclusief de tijdsplanning, waarbij de schakels overeenkomen met rustpunten. Daarmee bestaat de keten uit een verzameling van afspraken over rustpunten ten aanzien van de inhoudelijke betekenis, de kwaliteit en de planning van een rustpunt. Traditioneel – vanuit de stovepipe gedachte – waren dergelijke afspraken vaak bilateraal. Dit komt omdat een proceseigenaar niet verder de keten in hoefde te kijken (wilde kijken) dan zijn eigen raakvlakken. Bij ketenregie zijn deze afspraken 'ketenbreed'. De afspraken zijn SMART (Specific, Measurable, Acceptable, Realistic, Timely) en zodanig dat de keten relatief vaak tot een happy einde komt als iedere proceseigenaar zich aan de afspraak houdt. Een unhappy ketenflow ontstaat indien één of meer proceseigenaren zich niet aan de afspraak kunnen houden (vooraf geconstateerd door de eigenaar zelf) of hebben gehouden (achteraf geconstateerd door een gebruiker). Een afspraak wordt geschonden indien een gerealiseerd rustpunt niet aan de afgesproken inhoudelijke betekenis, kwaliteit en/of planning voldoet.

11.4 Borging (archivering)

Alle hierboven genoemde bedrijfsresultaten op het gebied van beleid, ontwerp, regie en uitvoering, dienen te worden geborgd of gearchiveerd. Figuur D-4 geeft een overzicht van de meeste in deze cursus besproken soorten bedrijfsresultaten.

Figuur D-4: Typische business objecten bij een statistisch proces



Deze typische bedrijfsresultaten zijn gegroepeerd naar de verschillende aspectgebieden. In architectuurterminologie worden bedrijfsresultaten en/of autonome onderdelen daarvan

ook wel aangeduid met *business objecten*; niet te verwarren met de real-world objecten uit figuur B-1 waarover we een statistiek willen maken.

Borgen (of archiveren) van de business objecten gaat verder dan bewaren. Borgen is bijvoorbeeld inclusief versie-, release- en toegangsbeheer.

- Zo kunnen er van rustpunten meerdere kwaliteitsversies zijn, die enerzijds uit elkaar gehouden moeten kunnen worden, maar anderzijds wel te relateren moeten zijn.
- Voor toegangsbeheer is het van belang verschillende rollen te onderscheiden. Bekend is de zogenaamde CRUD-matrix (Create, Read, Update, Delete) waarin aangegeven is welke rol wat mag doen.

Evenals de real-world objecten hebben business objecten eigenschappen. Ten behoeve van de borging kunnen bijvoorbeeld de volgende eigenschappen worden toegekend: (unieke) naam, versienummer, detentiedatum, eigenaar, gebruiker, beheerder.

Ook de 'omvang' in termen van aantallen bytes is een interessante eigenschap. Voor een statistische dataset waarbij het logische datamodel uit één objecttype bestaat met K variabelen, kan als vuistregel voor de opslag in een rechthoekig ASCII-formaat de omvang worden benaderd door

$$N\sum_{k=1}^{K} v_k$$

waarbij v_k het aantal benodigde posities voor variabele k is, $k = 1...K$, en N het aantal benodigde records. Uiteraard hangt deze benadering af voor het gekozen fysieke datamodel, zie paragraaf 3.3.

Een bestand met 20 miljoen eenheden (records) met 20 variabelen per eenheid en gemiddeld 2 posities per variabele zal ongeveer $20 \times 10^6 \times 20 \times 2 = 800 \times 10^6 =$ 0.8 GB groot zijn.

12. Context van het CBS (waarom-vraag)

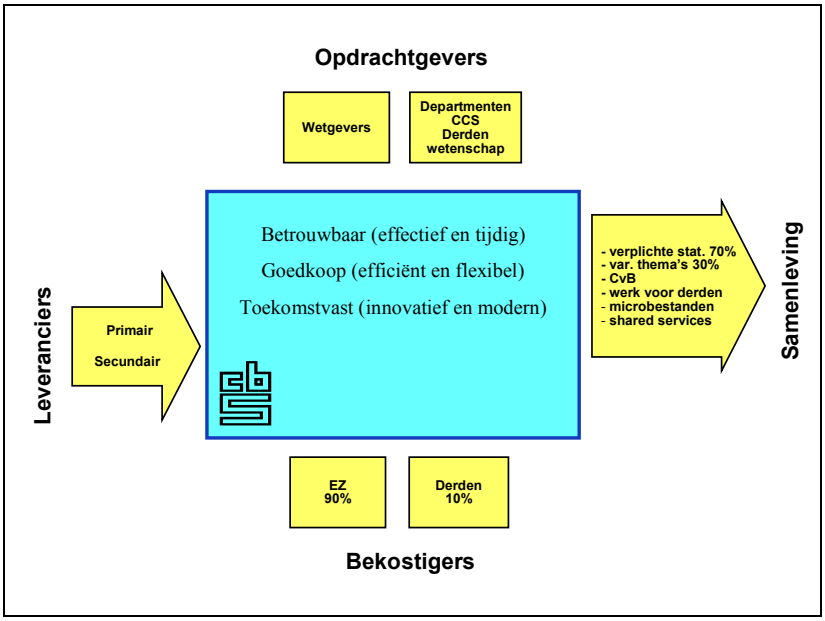
Uit de missie van het CBS mogen we concluderen dat de samenleving betrouwbare en samenhangende informatie wil, die inspeelt op haar behoeften (liefst natuurlijk tegen zo laag mogelijke kosten). Daartoe is het CBS ingesteld. Het vervullen van deze missie is daarmee de waarom-vraag en het bestaansrecht van het CBS.

Bij het vervullen van de missie zijn vanuit dezelfde samenleving vaak meerdere partijen betrokken met verschillende, soms tegenstrijdige, belangen. In figuur D-5 zijn een aantal van die partijen in beeld gebracht (overgenomen uit Bredero en Dekker, 2006).

Links staan de leveranciers van de benodigde statistische data, waarbij een onderscheid is gemaakt tussen primaire en secundaire waarneming. Rechts staan de afnemers van de statistische informatie. Opmerkelijk is dat ongeveer 70% van de totale statistiekproductie

van het CBS verplichte statistieken betreft waarover kennelijk beperkte marktvrijheid bestaat. Het CBS wordt voor een groot deel gefinancierd door Economisch Zaken (EZ) en heeft ten slotte te maken met wetgeving (denk aan bijvoorbeeld de CBS-wet). Midden in de figuur is het CBS als een black box uitgebeeld met een aantal 'ideale' kenmerken. Deze kenmerken komen deels voort uit opgelegde kaders en richtlijnen van buiten het CBS en deels vanuit de eigen CBS-ambitie.

Figuur D-5: Context van het CBS



Voor de komende jaren heeft het CBS de ambitie om de operational excellence, het productinnovatief vermogen en de klantgerichtheid te vergroten:

- Efficiëntie van het proces verhogen,
- Kwaliteit van product en proces verhogen,
  a. Betrouwbaarheid,
  b. Reproduceerbaarheid,
  c. Beheersbaarheid/ controleerbaarheid,
  d. Veiligheid,
  e. Tijdigheid,
- Verlagen ICT ontwikkel- en beheerkosten,
- Flexibiliteit (diversificatie en snelheid statistiekontwikkeling),
  a. Snelheid en kosten van statistiekontwikkeling,
  b. Snelheid en kosten van onderhoud productieproces/middelen,
  c. Veranderende aanleveringen m.b.t. primaire en secundaire bronnen,
- Productinnovatie; welke nieuwe producten voor welke klanten.

13. Procesinrichting (hoe-vraag)

Gegeven de missie van het CBS (de waarom-vraag) en haar ambitie voor de komende jaren zijn er vele mogelijkheden waarop het CBS haar vele processen kan inrichten. Belangrijke factoren die de inrichting van een proces bepalen zijn algemene kaders en richtlijnen (beleid, code of practice, architectuur), beschikbare middelen (zoals budget, beschikbare data, methodologische en technologische ontwikkeling), de huidige situatie met haar knelpunten (as-is), de gewenste ambitie (to-be), etc.

Idealiter wordt bij een (her)inrichting eerst in beeld gebracht wat de huidige situatie is met haar knelpunten (as-is), wordt vervolgens de gewenste situatie geschetst (to-be) die voldoet aan de algemene kaders en richtlijnen, en wordt ten slotte een migratiepad voorgesteld met tussenstappen die gegeven de beschikbare middelen haalbaar zijn. Een dergelijke analyse wordt in deze cursus een business analyse genoemd. Een business analyse kan COPAFIJTH-breed[13] wordt uitgevoerd, maar kan ook beperkt blijven tot enkele deelaspecten.

Deze paragraaf zal geen verhandeling geven over business analyses, maar zal in gaan op de kaders en richtlijnen vanuit de architectuur ten aanzien van de gewenste situatie.

13.1 Conceptuele architectuur principes

In deel B en C zijn de conceptuele aspecten en begrippen beschreven die idealiter onderdeel uitmaken van een statistisch product respektievelijk statistisch proces. Echter, bij veel huidige processen en producten ontbreken er aspecten – expliciet vastgestelde kwaliteitseisen of variabelendefinities zijn daar voorbeelden van – of zijn de aspecten onderling verstrengeld.

Ter bevordering van

- de relevantie, samenhang en betrouwbaarheid van de (uit te wisselen) producten,
- de effectiviteit, efficiëntie, transparantie, flexibiliteit en reproduceerbaarheid van de lokale processen, alsook
- de transparantie en efficiëntie van de gehele keten,

zijn vanuit de business architectuur zogenaamde architectuur principes geformuleerd, die als kader en/of als richtlijn worden meegegeven bij de (her)inrichting van een proces.

Deels hebben deze principes betrekking op de conceptuele en veelal abstracte aspecten en begrippen uit deel B en C. Dit zijn de zogenaamde conceptuele architectuur principes. Deels hebben de principes betrekking op de inzet van concrete business diensten en business middelen die het CBS ter beschikking stelt. Dit zijn de zogenaamde logische principes.

---

[13] De afkorting COPAFIJTH staat voor Communicatie (soms ook Cultuur), Organisatie, Personeel, Administratie, Financiën, Informatie, Juridisch, Technologie en Huisvesting.

Een voorbeeld van een business dienst is het Data Service Centrum (DSC). Een voorbeeld van een bedrijfsmiddel is de eenhedenbase (EHB). Merk op dat de beschikbaarheid en inzet van business diensten of van business middelen per statistisch bureau kunnen verschillen, maar dat de (abstracte) aspecten en begrippen ten aanzien van het statistische product en proces bij ieder statistisch bureau min of meer hetzelfde zijn.

Conceptuele architectuur principes worden in deze cursus opgevat als korte maar krachtige statements met toelichtingen, die enerzijds de essenties van deel B en C samenvatten en anderzijds als kader of richtlijn dienen om deze essenties bij een herontwerp na te leven (comply) dan wel afwijkende keuzes te verantwoorden (explain).

Tabel D-6 geeft de conceptuele architectuur principes. In de eerste kolom staan de principes in de vorm van statements, in de tweede kolom staat de (meest belangrijke) reden waarom dit principe nageleefd zou moeten worden en in de derde kolom staat een deelparagraaf waarin dit principe aan de orde komt.

Figuur D-6: Conceptuele business principes van het CBS

Statement	Reden	Korte toelichting
Er zijn geen data zonder metadata	Relevantie product	Het betreft hier de beschrijvende metadata; zie paragraaf 3.1 en 3.5
Eerst ontwerp dan uitvoering	Effectiviteit en repro-duceerbaarheid proces	Het betreft hier de voorschrijvende metadata; zie paragraaf 9.4
Ontwerp is klantgericht (en kostenbewust)	Relevantie product	Het betreft vooral externe klanten, maar ook interne; zie o.a. paragraaf 3.1 en 5.2
Ontwerp is gericht op maximaal hergebruik	Efficiëntie proces	Het betreft hier hergebruik van definities, rustpunten, generieke oplossingen, etc; zie o.a. paragraaf 5.2. en 9.8
Regie vindt plaats op basis van expliciete kwaliteitseisen	Transparantie keten	Dit betreft zowel keten- als procesregie m.b.t. meerdere kwaliteitsdimensies (samenhang, betrouwbaarheid, tijdigheid)
Opgeleverde data zijn in rust	Transparantie keten	zie paragraaf 5.2
Opgeleverde data met metadata worden uitgewisseld via koppelvlakken	Efficiëntie keten	zie paragraaf 5.4 en 6.1 – 6.3
De leverancier brengt, de klant haalt	Efficiëntie keten	Het betreft hier het brengen/halen van rustpunten naar/van koppelvlakken; zie paragaaf 5.3
Kwaliteitseisen zijn vertaald naar relevante, meetbare en normeerbare kwaliteitsindicatoren	Efficiëntie en transparantie proces	Het betreft hier een praktische operationa-lisering van kwaliteitseisen; zie paragraaf 5.5

De lijst met principes is niet compleet. Zo worden de flexibiliteit van het proces en de betrouwbaarheid en samenhang van het product niet geraakt. Sommige principes helpen wel – het principe *opgeleverde data zijn in rust* draagt bij tot samenhang – maar raken niet kern. In de volgende paragraaf komen aanvullende logische principes aan de orde.

13.2 Logische architectuur principes

Ook logische architectuur principes vatten we op als korte maar krachtige statements met toelichtingen. Deze statements hebben betrekking op enerzijds het gebruik van de door het CBS onderkende generieke proces diensten en anderzijds door het CBS onderkende typische (business) middelen.

Figuur D-7 geeft een aantal logische principes. Ook deze lijst is niet compleet en, in tegenstelling tot de lijst met conceptuele principes, niet formeel vastgesteld.

De eerste twee statements hebben betrekking op het gebruik van de generieke proces diensten. Voorwaarde daartoe is dat deze zijn ingericht. De overige vier statements hebben betrekking op een soort bedrijfsmiddel. Voorwaarde is dat deze bedrijfsmiddelen beschikbaar zijn.

Figuur D-6: Logische business principes van het CBS

Statement	Reden	Korte toelichting
Waarnemen is uitbesteed aan Dienst Dataverzameling (DV)	Efficiëntie proces	Het betreft hier de dienstenportfolio van waarnemen, zie paragraaf ???
Borging van rustpunten is uitbesteed aan het Data Service Centrum (DSC)	Efficiëntie keten	Het betreft hier de dienstenportfolio van DSC, zie paragraaf ???
Eenhedenbases zijn leidend bij de populatie-afbakening	Samenhang product	Het betreft hier de eenhedenbases voor bedrijven en personen, zie paragaaf ???
Classificatiebases zijn leidend bij het ontwerp van (classificerende) variabelen	Samenhang product	Het betreft hier de classificatieserver als CBS-brede lijst met well defined concept, zie paragraaf ??? en paragraaf 5.2
Methodovoorschriften zijn gebaseerd op valide methodologie	Betrouwbaarheid product	Het betreft hier de statistische methoden in de methodenreeks
ConGO-data zijn leidend bij de Keten Economische Statistieken	Samenhang product	Het betreft hier de statistische gegevens van *G*rote *O*ndernemingen, die ten aanzien van een divers aantal variabelen *Con*sistent zijn

Merk op dat het beschikbaar stellen van dergelijke bedrijfmiddelen ook weer kan worden opgevat als een soort (generieke) business dienst[14]. Deze diensten worden echter niet geleverd op verzoek van één specifieke klant (pull), maar worden als resultaat algemeen beschikbaar gesteld (push).

Merk verder op dat logische principes veel veranderlijker (kunnen) zijn dan conceptuele principes. Dat komt omdat conceptuele principes zich concentreren op de wat-vraag, i.e. de conceptuele aspecten van een statistisch product en proces, terwijl logische principes zich bezig houden met de hoe-vraag, i.e. algemene inrichtingskeuzes op CBS niveau.

Zo kan het CBS voor zichzelf beslissen het generieke dienstenpakket in te krimpen dan wel uit te breiden. Voor de hand liggende uitbreidingen van generieke procesdiensten zijn diensten op het gebied van koppelen, statistische beveiliging en distributie.

13.3 Generieke business diensten

In deze cursus vatten wij een generieke business dienst op als de levering van een (business) resultaat welke overeen gekomen is tussen twee partijen, namelijk de partij die het resultaat afneemt en de partij de het resultaat levert.

[14] Feitelijk kan het beschikbaar stellen van willekeurig rustpunt via een koppelvlak (en dus DSC) worden opgevat als een dienst met een aanbieder (eigenaar) en één of meer klanten (afnemers).

De partij die het resultaat levert wordt een Generieke ProcesDienst (GPD) genoemd, indien de dienst aan de volgende voorwaarden voldoet:

- De dienst functioneert autonoom (zelfstandig),
- De dienst heeft een vaste interface (contactpunt) voor haar afnemers,
- De dienst kent meerdere afnemers,
- De dienst onderhoudt een catalogus waarin staat welke soort diensten worden aangeboden, eventueel met meerdere servicelevels.

Met het laatste punt, i.e. de catalogus, wordt kenbaar gemaakt wat de scope van de GPD is. Refererend naar figuur D-1 kan deze scope afgebakend zijn naar één of meer aspectgebieden (ontwerp, regie, uitvoering en beheer) ten behoeve van één of meer business domeinen (waarnemen, verwerken en beveiligen, en informatieontwikkeling en publicatie), maar dat hoeft niet.

Momenteel kent het CBS twee Generieke Procesdiensten, namelijk Dataverzamelen (DV) en Data Service centrum (DSC).

13.3.1 Dienst DataVerzameling (DV)

Voor het domein waarnemen is een generieke procesdienst ingericht; dienst DataVerzameling (**DV**). Of deze dienst het complete waarneemdomein afdekt hangt af van diverse – veelal pragmatische – factoren. Momenteel bevat de dienstenportfolio van DV de volgende onderwerpen:

- Vragenlijstontwerp
- Steekproefontwerp
- Ontwerpen benaderstrategie
- Verzamelen data (via CATI / CAPI / CAWI / PAPI) op basis van deze ontwerpen inclusief mixed mode variant
- Secundaire waarneming – logistiek
- Monitoren ontvangst, appelleren en rappelleren (administratieve respons)
- Geautomatiseerde technische controles op ontvangen data
- Contactinformatie in één CRM applicatie
- Continue voortgangsinformatie door real-time raadpleegbare planningstool
- Levering data via Data Service Center (DSC)

Met afnemers worden in een zogenaamde waarneemopdracht afspraken vastgelegd welke diensten uit deze portfolio tegen welke kwaliteitseisen, kosten, tijdsplanning, etc. worden afgenomen.

In een ideale dienstenomgeving zullen vooral afspraken worden gemaakt over het productontwerp van de waarneemvraag, zie paragraaf 6.1, waarbij het aan DV wordt overgelaten *hoe* deze vraag wordt geoperationaliseerd en gerealiseerd. Het huidige dienstenportfolio laat echter vooralsnog diensten zien over de operationalisatie

(vragenlijstontwerp, steekproefontwerp, ontwerp benaderstrategie) en de realisatie (verzamelen data, monitoren, appelleren en rappelleren, technisch controleren, registreren contactinformatie en opleveren via DSC).

In hoeverre DV zal kunnen uitgroeien tot een generieke procesdienst waarin de wat-vraag centraal staat en niet de hoe-vraag, hangt sterk af van (het vertrouwen in) de kennis en kunde van DV om deze hoe-vraag correct te vertalen naar de wat-vraag.

13.3.2 Data Service Centrum (DSC)

Voor de borging van de rustpunten en de bijbehorende beschrijvende metadata, zie paragraaf 3.1, 3.2 en 3.3, is eveneens een generieke procesdienst ingericht; het Data Service Centrum (DSC). Momenteel bevat de dienstenportfolio van het DSC de volgende hoofdonderwerpen:

- Databeheer
- Metadatabeheer
- Beveiliging van bestanden
- Catalogusfunctie (over de rustpunten)
- Versiebeheer
- Autorisatie

Idealiter accepteert het DSC rustpunten met zowel complexe als enkelvoudige logische datamodellen, het brengen van deze rustpunten en halen van voorgedefinieerde selecties hieruit (zie ook paragraaf 3.3, 5.2 en 5.3).

Vooralsnog (1 januari 2009) is het alleen mogelijk om rustpunten met een enkelvoudige fysieke structuur naar DSC te brengen (de zogenaamde rechthoekige bestanden). Dit betekent dat rustpunten met een complex logisch datamodel eerst ‘platgeslagen moeten worden. Refererend naar paragraaf 3.3 kan dat vaak op verschillende manieren. Verder is het vooralsnog niet mogelijk (voorgedefinieerde) selecties uit deze platgeslagen rustpunten te halen. De platgeslagen rustpunten kunnen alleen als een geheel worden opgehaald.

Merk op dat in DSC-terminologie een logisch datamodel wordt beschreven aan de hand van het DataBronOntwerp (DBO). In Gelsema (2008) is het metadatamodel voor de borging van microdata in DSC beschreven.

13.4 Eenheden- en classificatiebases

Er zijn twee soorten van statistische data en metadata welke als ‘bedrijfsmiddel’ speciale aandacht verdienen, namelijk eenhedenbases en classificatiebases. Logisch gezien bestaat een eenhedenbase uit een verzameling van identificerende nummers (sleutels), coderingen, of beschrijvingen welke louter refereren naar real-world objecten. Classificatiebases bestaan logisch gezien uit verzamelingen van nummers, coderingen of beschrijvingen welke louter refereren naar eigenschappen.

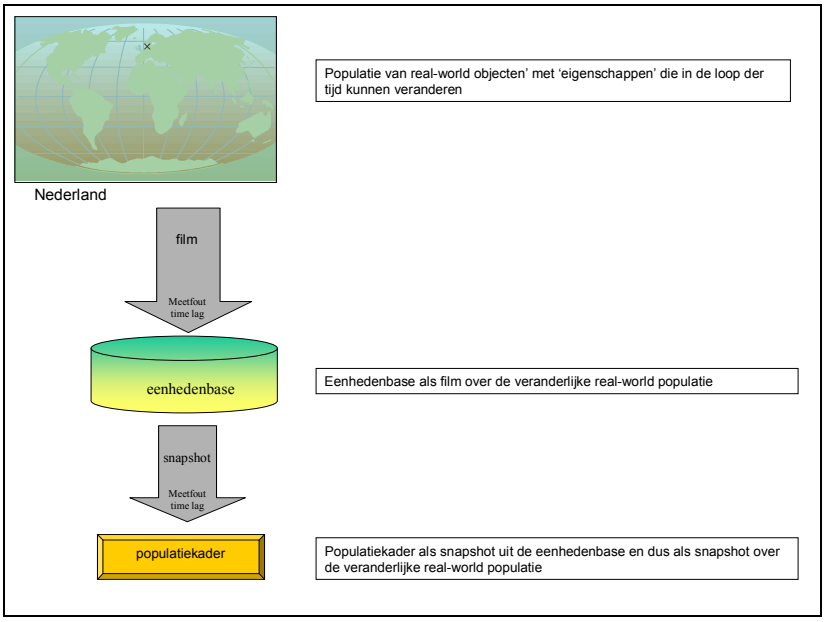
13.4.1 Eenhedenbases

We beschouwen alleen enkelvoudige eenhedenbases. Een ***enkelvoudige eenhedenbase*** kan worden opgevat als een verzameling van identificerende nummers (sleutels), coderingen of beschrijvingen welke verwijzen naar een ‘universum’ van real-world objecten. Een dergelijk ‘universum’ kan worden gedefinieerd aan de hand van een één of meer eigenschappen; elk real-world object dat deze eigenschappen bezit, behoort tot dit ‘universum’. Uiteraard kunnen real-world objecten met hun eigenschappen in de loop der tijd veranderen (geboorte, sterfte, immigratie, emigratie, fusie, splitsing, faillissement, huwelijk, scheiding, etc), wat betekent dat de samenstelling van het ‘universum’ in de loop der tijd kan veranderen. Deze zal over het algemeen groeien.

Het is de vraag of een eenhedenbase uit louter identificerende nummers moet bestaan. Het antwoord is ‘nee’. Om uit een eenhedenbase doel- of waarneempopulaties te kunnen selecteren met betrekking tot een specifieke periode, is ook een beperkt aantal attribuutvariabelen nodig, zoals geboortedatum en sterftedatum, waarmee de doel- en/of waarneempopulatie gedefinieerd is. Aan de hand van deze variabelen moet het mogelijk zijn om voor iedere gewenste periode een afbeelding van de gewenste populatie te selecteren. Een dergelijke subset uit de eenhedenbase noemen we een ***populatiekader***.

Het belangrijkste verschil tussen een eenhedenbase en een populatiekader is dat de eerstgenoemde opgevat kan worden als een film van real-world objecten, terwijl een populatiekader opgevat kan worden als een snapshot, zie figuur D-7

Figuur D-7: eenhedenbase versus populatiekader



Een voorbeeld van een enkelvoudige eenhedenbase is die van personen. Deze eenhedenbase refereert naar een ‘universum’ welke is gedefinieerd aan de hand van de eigenschap ‘is een persoon’ en verder afgebakend met de eigenschap ‘is ingeschreven in de Gemeentelijke Basis Administratie (GBA)’. Vanwege de statistische beveiliging zijn de

identificerende sleutels zoals de sofi-nummers (BSN-nummers) gepaard met zogenaamde betekenisloze RIN-nummers (Random Identifying Number).

Een voorbeeld van een complexe eenhedenbase is de EHB, zie Camstra en Renssen (2009). In deze eenhedenbase zitten meerdere objecttypen, waaronder de bedrijfseenheid.

Het zijn de ontworpen populatiekaders die via een koppelvlak (microbase) CBS-breed ter beschikking worden gesteld. De onderliggende eenhedenbases zorgen ervoor dat deze populatiekaders onderling consistent zijn, i.e. per versiemoment aan de zogenaamde 1-cijfergedachte voldoen. Op basis van (integrale) structuurtellingen en mutaties daarop wordt een eenhedenbase in de lokale verwerkingsomgeving voortdurend bijgewerkt. Uit de eenhedenbase kunnen op gezette tijden (wekelijks, maandelijks,..) door de eigenaar selecties worden gemaakt en als versies van populatiekaders (rustpunten) naar een koppelvlak worden gebracht. Welke versies en met welke frequentie is een kwestie van ontwerpen waarbij klantwensen in beeld komen.

13.4.2 Classificatiebases

Een classificatiebase is een verzameling van classificaties. We onderscheiden enkelvoudige en hiërarchische classificaties. Een ***enkelvoudige classificatie*** refereert naar een set van eigenschappen aan de hand waarvan een populatie van real-world objecten verdeeld kan worden in disjuncte, i.e. uitputtende en elkaar niet-overlappende, deelpopulaties (klassen). Indien deze classificaties afgebeeld zijn aan de hand van nummers, coderingen of beschrijvingen dan bestaat een enkelvoudige classificatie uit een set van nummers, coderingen of beschrijvingen. Enkelvoudige classificaties worden gebruikt om het waardedomein van een classificerende attribuutvariabele aan te geven of om disjuncte deelpopulaties te definiëren.

Naast enkelvoudige classificaties bestaat een classificatiebase uit hiërarchische classificaties. Een hiërarchische classificatie is een geneste rij van enkelvoudige classificaties, waarbij een enkelvoudig classificatie A genest is in een enkelvoudige classificatie B indien de set van eigenschappen die met A corresponderen een verfijning is van de set van eigenschappen die met B corresponderen. Bijvoorbeeld, beschouw gemeentecodes en provinciecodes als twee enkelvoudige classificaties. Dan zijn de ‘gemeentecodes’ genest in de ‘provinciecodes’.

13.4.3 Rol in het statistische proces

Zowel de classificatiebases als de eenhedenbases cq populatiekaders spelen op meerdere plekken in het statistische proces een belangrijke rol om statistische gegevens op elkaar af te stemmen (coördinatie).

- Eenhedenbases spelen een rol om populatieafbakeningen en begripsomschrijvingen op elkaar af te stemmen (hergebruik van populatieafbakeningen)
- Classificatiebases spelen een rol om classificerende variabelen op elkaar af te stemmen (hergebruik van classificaties).

Daarnaast spelen eenhedenbases cq populatiekaders een rol om de kwaliteit van de dekking van een dataset te meten en daarvoor eventueel te corrigeren.

Deel E: Herontwerpen onder Architectuur

14. (Her)ontwerp van statistische processen

Vanwege technologische en methodologische vooruitgang, veranderende wetgeving, kostenreductie, kwaliteitsverbeteringen, etc. zijn (grootschalige) herontwerpen van statistische processen van tijd tot tijd nodig. Zonder een uitputtende opsomming te (willen) geven bestaat een uitgebreide vorm van een herontwerp uit:

- het ontwerpen van de rustpunten die opgeleverd worden (productvoorschriften),
- het ontwerpen van (selecties) uit de benodigde rustpunten (productvoorschriften),
- het ontwerpen van de methodevoorschriften,
- het ontwerpen van de processen
    - procesvoorschriften
    - planningsvoorschriften,
- het implementeren van de voorschriften in (geautomatiseerde) systemen,
- het in beheer nemen van de uit het herontwerp resulterende artefacten (business objecten, zie paragraaf 11.4).

Vele (kleinschalige) herontwerpen op het CBS beperken zich tot de laatste twee onderdelen, i.e. het implementeren van oude processen in nieuwe geautomatiseerde systemen, zonder dat de in- en outputproducten, de methodenvoorschriften en of de processen zelf ingrijpend worden veranderd. Grootschalige herontwerpen gaan verder. Bij dergelijke herontwerpen is ruimte om opnieuw de methodologie te bekijken of om zelfs opnieuw met ‘de klant’ om tafel te zitten om het outputproduct en de kwaliteitseisen daaraan ter discussie te stellen.

Het herontwerpen van een statistisch proces kan vaak zelf weer worden opgevat als een proces, waarbij belangen van verschillende betrokkenen tegen elkaar afgewogen moeten worden. Er is vaak een discrepantie tussen wat aan de outputkant wenselijk is en wat aan de inputkant wordt geboden. Om dit gat te overbruggen is er is vaak weer een discussiepunt over een nieuw gat, namelijk het gat tussen wat theoretisch wenselijke methodologie en technologie is en wat praktisch haalbare methodologie en technologie is. Ook bij het ontwerpen van de procesmodellen kan onenigheid ontstaan over welk organisatieonderdeel over welk deel van het statistische proces verantwoording krijgt, etc. Het in beschouwing nemen van al deze ontwerpaspecten, waarbij dus niet een enkel aspect van het statistische proces wordt geoptimaliseerd, maar dat op het geheel wordt gelet, noemen we ‘total survey design’.

14.1 Verantwoording en beschrijving van herontwerpen

Wat de scope van een herontwerp ook is, de meeste herontwerpen worden vanwege het eindige karakter ervan in projectvorm gedaan.

Herontwerpen onder architectuur houdt onder meer in dat gedurende het project een aantal documenten tot stand komen waarin de gemaakte ontwerpkeuzes worden beschreven en verantwoord. We noemen de volgende documenten:

- Het Methodologisch Advies Document (MAD) beschrijft en verantwoordt de statistische methodologie en relateert deze aan de statistische methoden die daaraan ten grondslag liggen.
- Het Business Architectuur Document (BAD) beschrijft de statistiekbehoefte van de afnemer en het gehele procesmodel (de productvoorschriften voor de in- en outputs, de methodevoorschriften, de procesvoorschriften en planningsvoorschriften) en relateert dit procesmodel aan de statistische methodologie die daaraan ten grondslag ligt.
- Het Software Architectuur Document (SAD) is afkomstig van RUP en beschrijft de geautomatiseerde oplossing voor het ontworpen procesmodel op basis van een vijftal views: use case view, logical view, process view, implementation view en deployment view. Hier moet ook tot uiting komen of (delen van) de oplossing gerealiseerd gaat worden door middel van Commercial off the Shelf (COTS), regelsturing of zelfbouw.

De idee is dat gedurende het herontwerp- en ontwikkelproces een aantal review-momenten is ingelast om bovengenoemde documenten te toetsen.

14.2 Aspecten van een BAD

Een BAD geeft een beschrijving van het gewenste statistische proces, waarbij de nadruk ligt op het gebruiken van de generieke procesdiensten en het consistent zijn met de CBS business architectuur. Het doel van een BAD is meervoudig.

- Sturend: gegeven doel en externe omstandigheden van een statistiek levert de BAD vanuit business perspectief een ontwerp (bouwtekening) voor het product en proces, waarmee de business haar doel kan behalen en waarmee vervolgens de IT-architect aan de slag kan gaan om tot een IT-oplossing te komen.
- Toetsend: aan de hand van de BAD vindt een toetsing plaats op de wijze van product- en procesontwerp door deze tegen de uitgangspunten van CBS brede business architectuur aan te houden.
- Beschrijvend: een BAD documenteert en maakt achteraf inzichtelijk de geïmplementeerde processen en daaruit voortvloeiende producten vanuit business perspectief, op basis waarvan later change management gevoerd kan worden.

Afhankelijk van de complexiteit van een herontwerp en de ambitie om het herontwerpproject onder architectuur te willen uitvoeren kan het detailniveau van een BAD verschillen.

Tabel E-1 geeft de onderwerpen die in een BAD aan de orde moeten komen. De eerste zes onderdelen volgen de ontwerpdomeinen van een statistiek conform de Business Architectuur, zie de gele wiebers in de bovenlaag van figuur D-1. Deze onderdelen zijn aan de orde gekomen in deel B en C.

Tabel E-1: Onderwerp van een BAD

	Onderdeel	Waarom in BAD
1	Bestaansrecht van de statistiek en haar belangrijkste afnemers	Geeft antwoord op de waarom-vraag; geeft kaders aan de wat- en hoe-vraag.
2	Reden (business case) van het project	Geeft antwoord op de waarom-vraag; geeft kaders aan de wat- en hoe-vraag.
3	Overzicht van de op te leveren rustpunten in termen van productvoorschriften.	Geeft antwoord op de wat-vraag; geeft kaders voor de hoe-vraag.
4	Overzicht van de benodigde rustpunten in termen van productvoorschriften.	Is onderdeel van de hoe-vraag.
5	Overzicht van de methodevoorschriften.	Is onderdeel van de hoe-vraag.
6	Overzicht van het proces (en de proces- en planningsvoorschriften).	Is onderdeel van de hoe-vraag.
7	Context van het proces.	Geeft externe partijen en dus factoren aan waarmee het proces te maken heeft
8	Verantwoording van de B&I-principes.	Comply or explain.

Het zevende onderdeel – de context van het proces – betreft de omgeving waarmee het proces te maken heeft, zie paragraaf 17. Deze omgeving kan zowel binnen als buiten het CBS liggen. Dit in tegenstelling tot de context van het CBS, zie paragraaf 12, dat per definitie alleen buiten het CBS ligt.

Het achtste onderdeel geeft aan in hoeverre een herontwerpproject in staat is geweest gehoor aan te geven aan de architectuur principes, zie paragraaf 13.

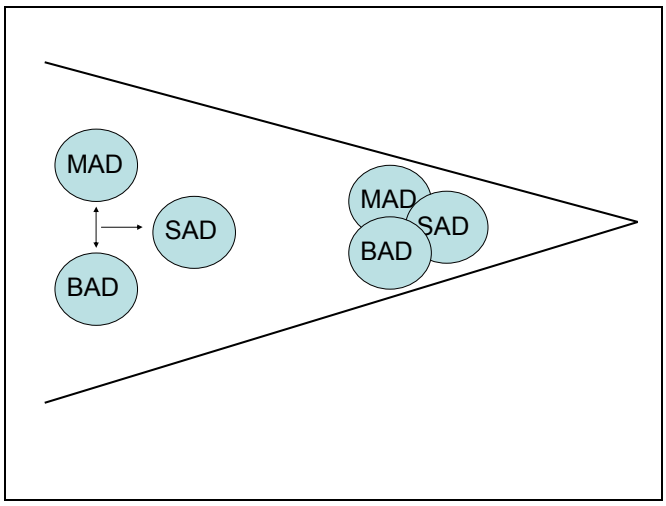
14.3 Herontwerp vanuit TSD-perspectief (Total Survey Design)

Idealiter vindt het (her)ontwerp van een statistisch verwerkingsproces outputgericht plaats, dat wil zeggen dat gegeven de statistische informatiebehoefte wordt vastgesteld

- de op te leveren rustpunten (productvoorschriften),
- de benodigde (selecties) uit rustpunten (productvoorschriften),
- de methodologie om het ‘gat’ tussen beschikbare rustpunten als inputs en de op te leveren rustpunten als gewenste outputs te dichten (methodevoorschriften).

De gewenste rustpunten en de daarvoor benodigde (selecties uit) rustpunten worden in de BAD beschreven. In de BAD worden ook de processen beschreven. De benodigde methodologie om het 'gat' te dichten wordt in de MAD beschreven. Zodra er (enige) helderheid is over de producten, methodologieën en processen kan een start worden gemaakt met de SAD.

Figuur E-1: het convergentieproces tussen MAD, BAD en SAD



Idealiter dienen bij een herontwerp de MAD's, BAD's en SAD's door de hele keten heen in samenhang opgesteld te worden, waarbij vanuit een soort Total Survey Design (TSD) perspectief de verschillende documenten voor de verschillende ketenprojecten tot stand zouden moeten komen. Echter, in de praktijk zullen in de beginfase van het herontwerp de MAD, BAD en SAD per project enigszins als geïsoleerde artefacten worden opgesteld, waarbij er zoals hierboven beschreven nog wel een één-tweetje is tussen MAD en BAD. Pas in een later stadium zullen de drie artefacten ketenbreed naar elkaar toe moeten groeien, waarbij er in alle documenten ongetwijfeld concessies moeten worden gedaan, zie figuur E-1.

Dat ketenbreed convergeren tussen BAD, MAD en SAD is een iteratief proces, waarbij vaak lastige keuzes moeten worden gemaakt, doordat verschillende belanghebbenden – de 'business'[15], ITS, methodologie en generieke diensten als DSC en DV – een rol spelen. Immers, bij een goed herontwerp vindt er een voortdurende afweging plaats tussen kosten en baten.

Daarbij zullen de baten vooral bij business liggen (de uiteindelijke statistische producten met bepaalde kwaliteit) en de kosten bij zowel de business als bij ITS, methodologie en de generieke diensten. Aan de kostenkant is het bovendien van belang een onderscheid te maken tussen (éénmalige) ontwerp- en ontwikkelkosten en (structurele) beheer- en productiekosten.

[15] Het gaat om het 'business-gevoel' om samen verantwoordelijk te zijn voor de hele keten.

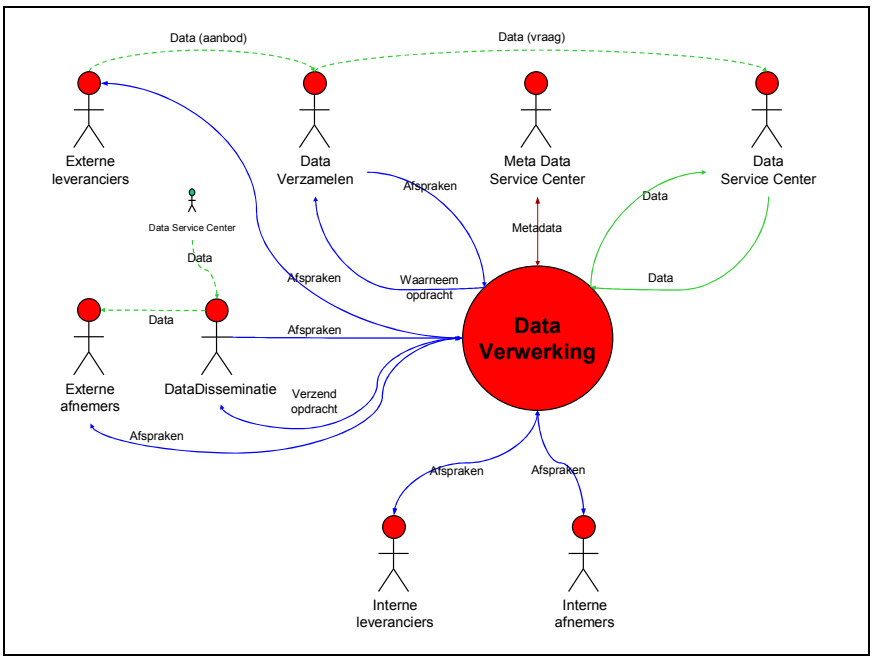
15. Business context van een verwerkingsproces

In figuur E-2 is een typische business context van een verwerkingsproces weergegeven. In deze figuur zijn de ‘poppetjes’ businessactoren buiten het verwerkingsproces waarmee de businessactoren binnen het verwerkingsproces (de grote rode bol) te maken hebben. Typische business actoren van buiten het proces zijn

- Externe en interne afnemers
- Externe en interne leveranciers
- Generieke procesdiensten
    - Data Service Centrum (en Meta Data Service Centrum)
    - Dienst DataVerzamelen

Merk hierbij dat de rol van afnemers en klanten als business actor van een andere aard is dan de rol van een generieke procesdienst. De positie van afnemers en klanten zal altijd in de context zijn van een verwerkingsproces zijn, terwijl de positie van een generieke procesdienst afhangt van algemene inrichtingskeuzes op CBS niveau. Zodra een generieke procesdienst ophoudt te bestaan, worden de processen die bij deze dienst horen overgenomen door de ‘rode bol’. Andersom, indien er een generieke procesdienst bijkomt, kunnen er processen worden overgenomen door deze dienst. Ter illustratie is in figuur E-2 een (nog) niet bestaande generieke procesdienst als business actor toegevoegd, namelijk Dienst DataDisseminatie (distributie).

Figuur E-2: Business context van een verwerkingsproces



De doorgetrokkken blauwe pijlen in figuur E-2 stellen afspraken voor die een proceseigenaar maakt met de business actoren. Deze afsprakenstromen voltrekken zich in de ontwerpfase.

De doorgetrokken groene pijlen stellen datastromen voor tussen enerzijds de businessactoren binnen het verwerkingsproces en anderzijds de businessactoren buiten het verwerkingsproces. De gestippelde groene pijlen stellen datastromen voor tussen de verschillende businessactoren buiten het verwerkingsproces. Strikt genomen horen deze datastromen niet thuis in een contextdiagram, maar voor de volledigheid zijn zij toch opgenomen. De datastromen voltrekken zich in de uitvoeringsfase.

Een business context zoals in figuur E-2 hoort thuis in een BAD – het zevende onderdeel uit tabel E-1 – waarbij de afspraken met zowel de generieke procesdiensten als met de specifieke klanten en leveranciers zijn ingevuld.

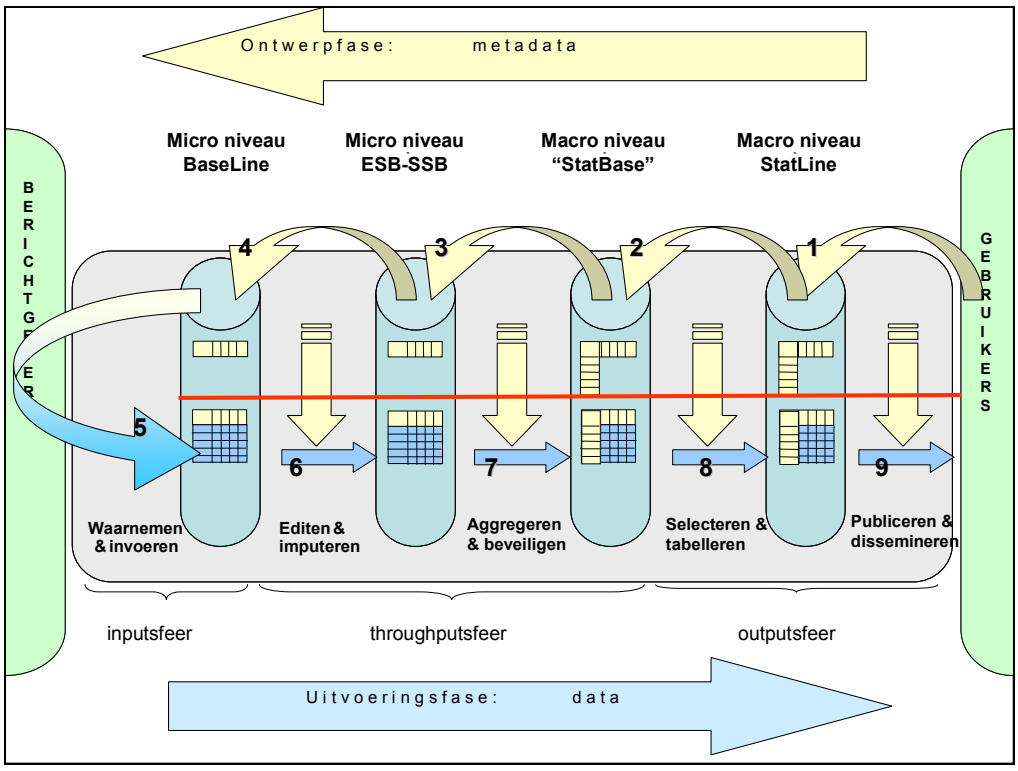
Deel F: Verwante Architectuur (Modellen)

16. Voorloper van de CBS-architectuur

Willeboordse (2005) en Willeboordse, Struijs en Renssen (2006) zijn belangrijke inspiratiebronnen geweest voor de CBS-architectuur zoals deze op basis van deel B en C is uiteengezet in Deel D. We gaan kort in op de visie van het statistische proces van toen.

Figuur F-1 geeft de logische fasen en momenten in het statistische proces, zoals deze in het 'pre-architectuur' tijdperk werden gedoceerd in Willeboordse (2005). De figuur laat zien dat er sprake is van een cyclus, waarin de gebruikers (externe klanten) het begin- en eindpunt vormen, en de berichtgevers (externe leveranciers, respondenten) het keerpunt. De cyclus laat een helder onderscheid zien tussen ontwerp (boven) en uitvoering (onder). In de CBS-architectuur zijn over de grondgedachte van de cyclus principes geformuleerd, die deze grondgedachte expliciteren. Dit zijn onder andere de principes 'er is geen data zonder metadata', 'eerst ontwerp en dan uitvoering' en 'ontwerp is klantgericht'.

Figuur F-1: Logische fasen en momenten in het statistische proces



In de figuur zijn de voorlopers van de koppelvlakken (Baseline, Microbase, Statbase en StatLine) herkenbaar. In de CBS-architectuur zijn deze voorlopers uitgewerkt tot de huidige koppelvlakken, die een belangrijke archieffunctie spelen voor zowel de data als de metadata. Als belangrijke toevoeging aan het concept *koppelvlak* is het concept *rustpunt* geïntroduceerd en zijn er principes als 'opgeleverde data zijn in rust' en

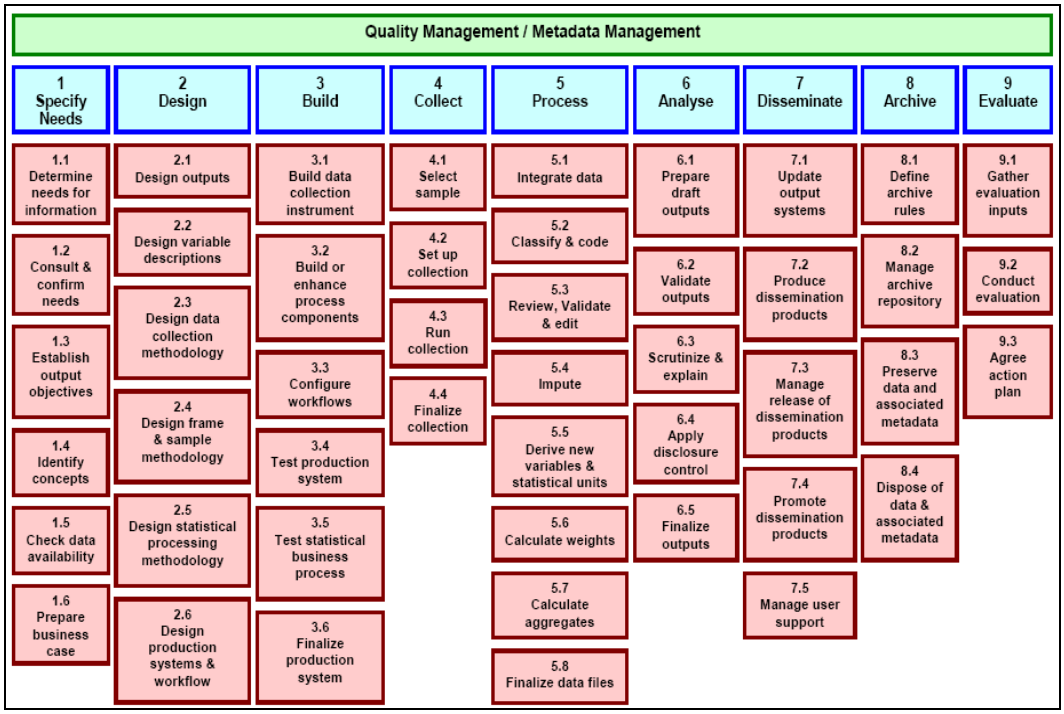
‘opgeleverde data met metadata (rustpunten dus) worden uitgewisseld via koppelvlakken’ waarbij ‘de leverancier brengt en de klant haalt’ geformuleerd.

Een gemis van figuur F-1 betreft de regiefunctie (zowel de proces- als ketenregie). Later, in de CBS-architectuur is deze functie onderkend als belangrijke schakel tussen ontwerp en uitvoering. Doordat statistische processen steeds meer als ketens worden geïmplementeerd groeit het belang van ketenregie op rustpunten, waarbij de rustpunten niet alleen de (tussenliggende) statistische producten zijn, maar tevens de schakels in de keten. De principes ‘regie vindt plaats op basis van expliciet vastgestelde kwaliteitseisen’ en ‘kwaliteitseisen zijn vertaald naar relevante, meetbare en normeerbare kwaliteitsindicatoren’ zijn geformuleerd ten behoeve van de regiefunctie.

17. Generic Statistical Business Process Model (GSBPM)

Een architectuur voor statistische processen die momenteel in de belangstelling staat is het Generic Statistical Business Process Model (GSBPM). In vergelijking met de CBS-architectuur moet GSBPM niet als een procesmodel worden opgevat, maar als een geordende classificatie van typische statistische activiteiten op onder meer het gebied van ontwerp, uitvoering en borging (archivering). Voor zover deze activiteiten betrekking hebben op het ontwerp- en uitvoeringsdomein, zijn zij vergelijkbaar met het ontwerpen en uitvoeren van de typische handelingen en standaard processtappen uit paragraaf 9.

Figuur F-2: Classificatie van typische statistische activiteiten conform het GSBPM-model



In de inleiding van deze cursus is aangegeven wat de toegevoegde waarde van architectuur is. GSBPM was oorspronkelijk ontwikkeld om statistische organisaties, zoals

het CBS, te helpen in hun *communicatie* over statistische processen met hun bijbehorende metadatasystemen door overeenstemming te bereiken over terminologie. Echter, gaandeweg is het doel van GSBPM verbreed. Zodra overeenstemming is over terminologie, kan de weg worden geopend naar bijvoorbeeld internationale samenwerking op het gebied van softwaredeling.

Het GSBPM-model is beschreven in Vale (2009) en bestaat uit vier niveau's:

- het gehele statistische proces,
- negen fasen in het statistische proces,
- deelfasen in ieder van de negen fasen,
- beschrijvingen van de deelfasen,

waarbij de negen fasen met hun deelfasen de classificatie vormen. In figuur A-2 is deze classificatie uit Vale (2009) overgenomen. Ieder deelfase, i.e. iedere typische statistische activiteit, heeft een aantal kenmerken

- Input(s)
- Output(s)
- Doel (toegevoegde waarde)
- Eigenaar
- Richtlijnen (zoals handleidingen)
- Triggers (enablers)
- Feedback loops.

Conform het model kunnen deze eigenschappen echter per statistisch proces verschillen en zijn daarom binnen GSBPM niet verder uitgewerkt.

Een vergelijking met de CBS-architectuur leert ons dat GSBPM evenals de CBS-architectuur richting 'bouwstenen' wenst te gaan voor het statistische proces. Er zijn redelijk wat bouwstenen (typische statistische activiteiten) door GSBPM geïdentificeerd, maar deze bouwstenen zijn niet uitgewerkt in termen van handelingen, standaard processtappen en standaardprocessen.

In GSBPM wordt Quality Management onderkend, zie ook figuur Figuur-2, dat sterk gerelateerd is met de evaluatiefase uit GSBPM. Het model achter dit Quality Management concept is de DEMING-cyclus (PLAN, DO, CHECK, ACT). De DEMING-cyclus en daarmee Quality management is van toepassing op iedere deelfase in GSBPM. Echter, evenals het concept van de 'bouwstenen' is Quality Management binnen GSBPM niet verder uitgewerkt. Het is vergelijkbaar met de procesregie in de CBS-architectuur.

GSBPM lijkt gericht te zijn op de klassieke stovepipe-achtige processen, omdat de gehele ketenproblematiek, inclusief het (intern) uitwisselen van data, nauwelijks of geen

aandacht krijgt. In de CBS-architectuur speelt de ketenproblematiek juist een prominente rol, waarvoor de concepten ‘rustpunt’ en ‘ketenregie’ zijn geïntroduceerd.

18. Algemene modelleringstalen

De cursus wordt afgesloten met een korte bespreking van een aantal algemene modelleertalen voor het beschrijven en vastleggen van bedrijfsprocessen c.q.bedrijfsinformatie.

18.1 Unified Modeling Language (UML)

Een voorbeeld van een dergelijk model is UML (Unified Modeling Language) waarin de logische datamodellen in bijvoorbeeld figuur B-3 en figuur C-13 zijn vastgelegd. UML is een standaard voor het specificeren van softwareproducten en is een essentieel onderdeel van RUP (Rational Unified Process). RUP is binnen het CBS een geaccepteerde standaard en dat verklaart waarom relatief veel van onze informatiemodellen in UML zijn vastgelegd.

UML lijkt vooral een geschikte taal om informatie over businessobjecten en relaties tussen deze objecten te modelleren, zie ook paragaaf 11.4. Refererend naar het IAF-framework (figuur A-1) kan UML worden gepositioneerd in de Informatiekolom.

Een nadeel van UML is dat het geen ‘natuurlijke’ taal is en dat veel business mensen deze taal moeilijk kunnen begrijpen.

18.2 Business Process Model and Notation (BPMN)

BPMN is een diagrammatische standaard voor het modelleren van bedrijfsprocessen, zie Nijssen en Le Cat (2009). Met BPMN kan worden aangegeven welke actoren welke processen, taken en beslissingen uitvoeren, alsmede welke actoren welke boodschappen opstellen en sturen naar andere actoren. Het is daarmee een communicatietaal voor bedrijfsprocessen en hun samenstellende elementen.

Refererend naar figuur A-1 en toegepast op het CBS is BPMN vooral geschikt om in de Businesskolom de statistische processen te modelleren op zowel de conceptuele laag, de logische laag als de fysieke laag.

18.3 Semantics of Business Vocubulary and Business Rules (SBVR)

SBVR is een op natuurlijke taal gebaseerde standaard voor het gestructureerd beschrijven van feiten, zie wederom Nijssen en Le Cat (2009). Refererend naar het IAF-framework van figuur A-1 speelt SBVR zich af in de Business- en Informatiekolom.

Toegepast op het CBS is zij vooral geschikt om het statistische product te beschrijven op zowel de conceptuele laag, de logische laag als de fysieke laag.

18.4 CogNIAM

CogNIAM (Cognition enhanced Natural language Information Analysis Method) is een methode welke BPMN en SBVR combineert door te onderkennen dat ook bedrijfsprocessen als feiten kunnen worden gemodelleerd, Nijssen en Le Cat (2009). Refererend naar het IAF-framework van figuur A-1 speelt CogNIAM zich af in de Business- en Informatiekolom.

Toegepast op het CBS is CogNIAM geschikt om statistische processen te beschrijven, inclusief de gehanteerde voorschriften, op zowel de conceptuele laag, de logische laag als de fysieke laag.

Geraadpleegde literatuur

Booleman, M, Buiten, G. Gelsema T, Lammertsma, A. en Renssen, R. (2010), Visie op Ketenregie, notitie van 21 januari 2010.

Bredero, R. en Dekker, W. (2006), projectgroep architectuur (2006), Programma Procap, ICT-Masterplan: CBS-architectuur, Context van Verandering, versie 2.0, CBS.

Camstra, A. en Renssen R.H. (2009), Eenhedenbase vanuit de Business Architectuur, Interne CBS-nota, Heerlen.

Daas, P. en Arends-Tòth, J. (2008), Kwaliteitsaspecten van Registers: het Begrippenkader, CBS, Heerlen.

Daas, P. en Beukenhorst, D. (2008), Databronnen van het CBS: Primaire en Secundaire bronnen, conceptversie 0.5.

Dekker, W. (2009), BI-Architectuur Outputdomein, Rapportage Fase II.

Delden van, A., Daas, P., Heerschap, N., Lammertsma, A., Nederpelt van, P. en Renssen R. (2008), Kader voor het Organiseren van Ketenregie, CBS, Voorburg.

Gelsema, T. (2008), Toelichting voor het Lezen van de Diagrammen uit het Metadatamodel DSC, notitie van 26 november 2008.

Gelsema, T en Windmeijer, D (2007), ICT-Masterplan: CBS-architectuur, Conceptuele Business Architectuur Metadata, versie 4.0, CBS.

Huigen, R., projectgroep architectuur (2006a), ICT-Masterplan: CBS-architectuur, Business- en Informatiemodel (BI001), versie 1.0, CBS.

Huigen, R.H. (2006b), Onderscheidende Architectuur Principes; Memo van Architectuur aan Doelgroep, notitie van 15 maart 2006.

ISO (2005) ISO/IEC 11179 – Metadata Registers (all parts), Geneva: International Organisation for Standardisation and International Electrotechnical Commision.

Kuijvenhoven, L. en Schouten, B. (2008), Kwaliteitsindicatoren voor de Hyperdimensie Data, Interne CBS-nota, Voorburg.

Lammertsma, A. (2008), Houtskoolschets Verwerking KS in Kwartaalrekeningen, Informele Gedachtenuitwisseling in pdf-formaat.

Loggen, R. (2008), Eindrapport Requirementsanalyse Statline Toekomst, versie 0.9.

Projectgroep afbakening SSB-kern en SSB-satellieten (2007), Afbakening van de SSB-kern en de SSB-satellieten, Concept Eindrapport (versie 1.1), Voorburg.

Projectgroep Architectuur, PROCAP, (2006), ICT-Masterplan: CBS-architectuur, Context van Verandering (CX001), versie 2.0, CBS.

Renssen, R.H. (2007), Ontwerp van een Transformatievoorschrift: van Pre-inputbase naar Inputbase, versie 1.0, CBS.

Renssen, R. (2007), DSC voor CBS-brede Uitwisseling Statistische Data. Interne CBS-nota, Heerlen.

Renssen, R. (2008), Ketenregie, KS met BTW als Voorbeeld, Interne CBS-nota, Heerlen.

Renssen, R. (2008), Ruggengraten; Eenhedenbases voor Statistische data, Interne CBS-nota, Heerlen.

Renssen, R. (2008), Producing Statistics from a Business Perspective, Statistics Netherlands, Heerlen.

Renssen, R. en Camstra A. (2008), Statistische beveiliging van persoonsgegevens voor intern gebruik, Heerlen.

Renssen, R. en Kruiskamp, P. (2008), Pilot KS met BTW-gegevens voor DSC, CBS, Heerlen.

Renssen, R., Morren, M., Camstra, A. en Gelsema T. (2009), Standaard Processen, Interne CBS-nota, Heerlen.

Renssen, R., Paulussen, R. Nauts, H. en Visser, W. (2008), BAD IIS (Inkomens Informatie Systeem), Interne CBS-nota, Heerlen.

Renssen, R., Puts, M. en Hellenbrand, R. (2008), BAD binnenvaart, Interne CBS-nota, Heerlen.

Sundgren, B., Androvitsaneas, C. and Thygesen, L. (2006), Towards an SDMX User Guide: Exchange of Statistical Data and Metadata between Different Systems, National and International, Organisation for Economic Co-orporation and Development, Geneva.

Sundgren, B., Thygesen, L. and Ward D. (2007), A model for Structuring of Statistical Data and Metadata to be Shared between diverse National and International statistical Systeems, second draft (work in progess to be presented at the OECD Expert group on SDMX).

Vale, S. (2009), Generic Statistical; Business Model, Joint UNECE/Eurostat/OECD, Work Session on Statistical Metadata (METIS), Version 4.0.

Van Delden, A en B. Braaksma (2007), Architectuur HEcS, Contourplaat: rustpunten, Interne CBS-nota, Centraal Bureau voor de Statistiek, Voorburg.

Van der Laar, R. (2008), Conceptuele typering van processtappen naar business functie, Interne CBS-nota in wording, Voorburg.

Van der Plas, J. en Sluis, W. (2006), DIGROS Registermodel, Interne CBS-nota, versie 1.0, CBS, Voorburg.

Vosmer, M., Van Berkel, K., Cuppen, M., Janssen, B., Nauts, H. en Renssen, R. (2004), Progriss Maquette Cardan, Interne CBS-nota, Heerlen.

Willeboordse A. (2005), Logische Grondslagen van de Beschrijvende Statistiek, Startcursus Methodologie, Module 1.

Willeboordse, A. (2008), Inleiding in de Methodenreeks en het Statistisch Proces, Interne CBS-nota, Voorburg.

Willeboordse, A., Struijs, P., and Renssen, R. (2006), On the Role of Metadata in Achieving Better Coherence of Official Statistics, CBS.

Willenborg, L. (2008), Pre- en Postcondities bij Standaard Processtappen, Interne CBS-nota in wording, Voorburg.

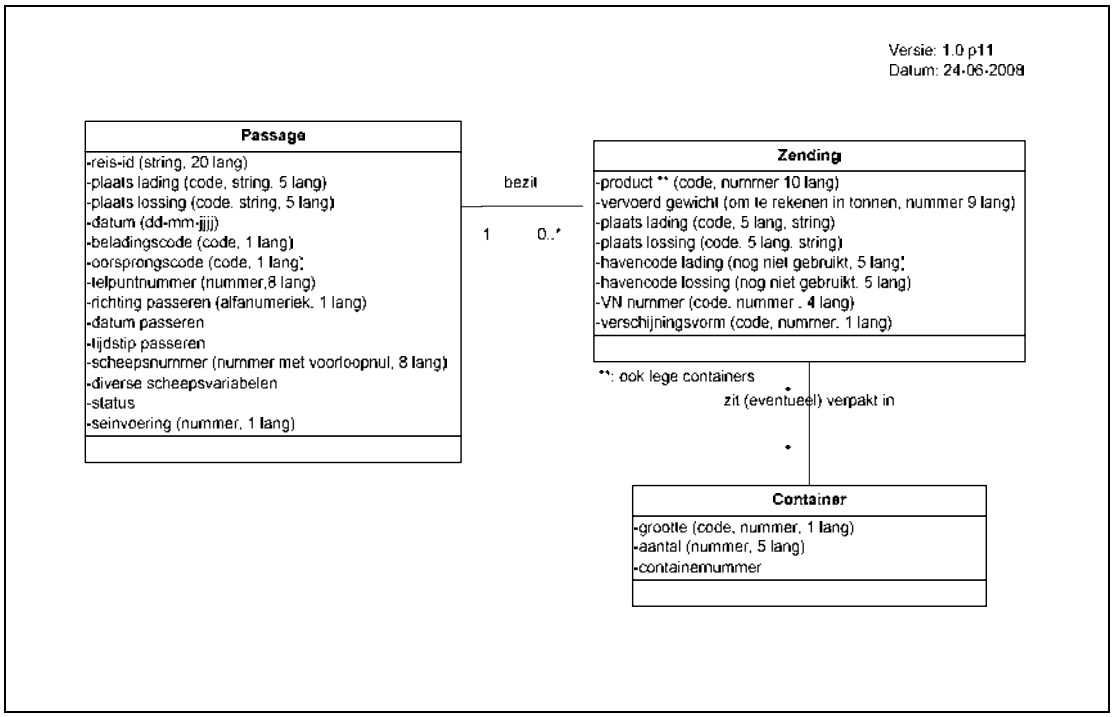
Windmeijer, D. (2006), ICT-Masterplan: CBS-architectuur, Logische Informatie Architectuur (BI002), versie 1.0, CBS.

Appendices

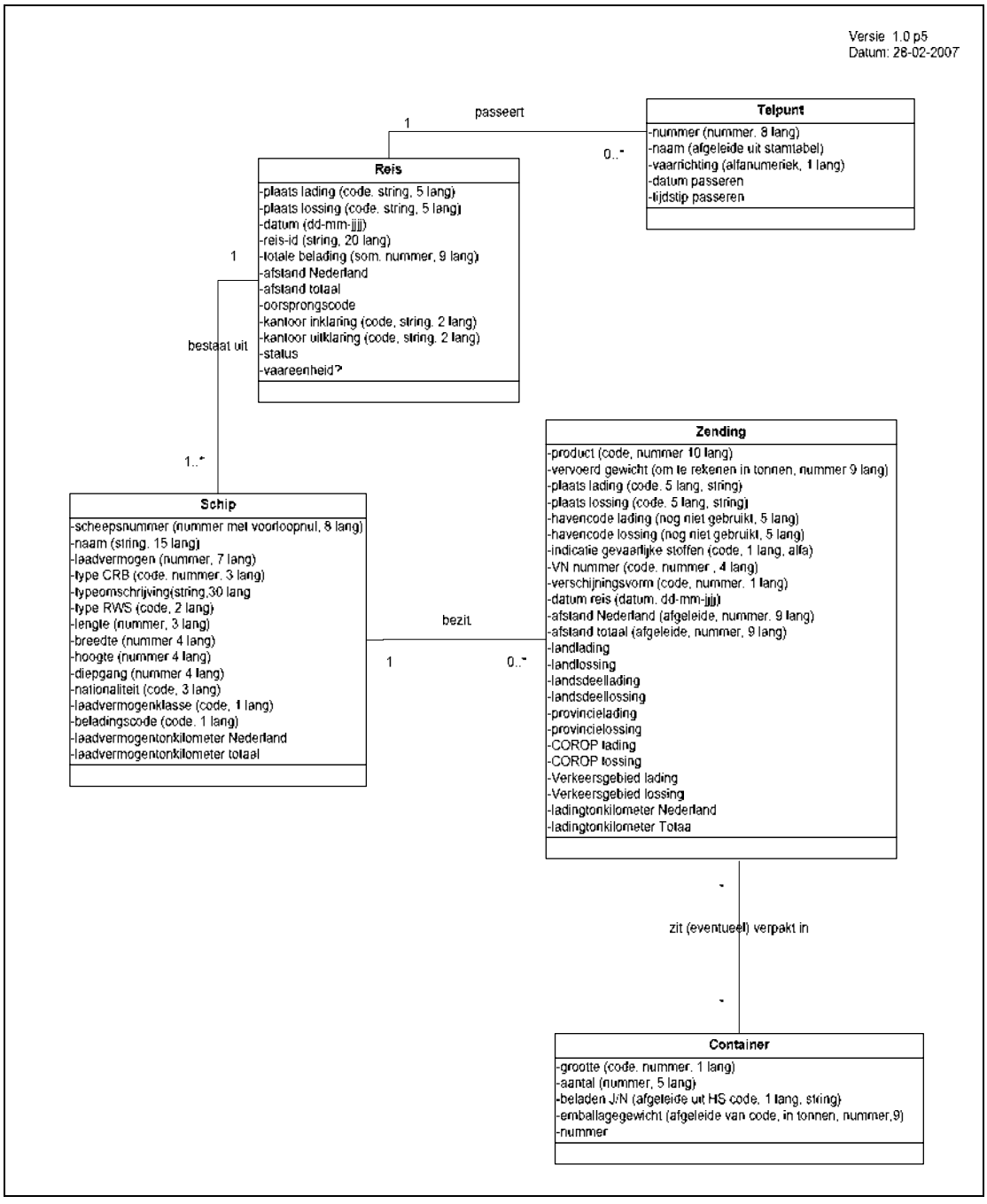
Appendix A: logisch datamodel leveringen Eurostat (tabellen)

Tabel	*Objecttype*	*Afbakening*	*Classificerende variabele*	*Kwantificerende variabele*
A1	zending	op nederlandse binnnenwateren	plaats lading plaats lossing goederensoort (product) verpakkingssoort (verschijningsvorm)	vervoerd gewicht lading tonkilometers (Ned)
B1	zending	op nederlandse binnnenwateren	plaats lading plaats lossing soort schip nationaliteit schip	vervoerd gewicht lading tonkilometers
B2	schip	op nederlandse binnnenwateren	soort vervoer beladingscode	aantal schepen afgelegde afstand
C1	container	op nederlandse binnnenwateren	plaats lading plaats lossing goederensoort containersoort beladingstoestand containers	vervoerd gewicht tonkilometers TEU TEU-km
D1	zending	op nederlandse binnnenwateren	kwartaal soort vervoer nationaliteit schip	vervoerd gewicht tonkilometers
D2	container	op nederlandse binnnenwateren	kwartaal soort vervoer nationaliteit schip beladingstoestand containers	vervoerd gewicht tonkilometers TEU TEU-km
E1	zending	op nederlandse binnnenwateren	soort vervoer goederensoort	vervoerd gewicht tonkilometers

Appendix B: logisch datamodel van een IVS-levering (microbestand)



Appendix C: logisch datamodel 'integraal' gaafgemaakt microbestand



Appendix D: variabeledefinities

Variabelen bij Zending	***Definitie***	***Domein***
Product	Codering soort vervoerde goederen	HS-code
vervoerd gewicht	Gewicht van de zending (in tonnen; 1000 kg)	numeriek, 2 decimalen
plaats lading	Plaats waar goed geladen is	UNLO-codering (pos. 1t/m 5)
plaats lossing	Plaats waar goed gelost is	UNLO-codering (pos. 1t/m 5)
havencode lading	Haven waar goed geladen is	UNLO-codering (pos. 6 t/m10)
havencode lossing	Haven waar goed gelost is	UNLO-codering (pos. 6 t/m10)
VN-nummer	Aanduiding van Verenigde Naties voor soort gevaarlijke stoffen	VN-codes: 0000 t/m 9999
Verschijningsvorm	Verpakkingstoestand van een zending	0 t/m 9, zie tabel
datum reis	passagedatum bij geselecteerd telpunt	Datum
afstand Nederland	afstand tussen plaats lading tot plaats lossing van een zending (in km), exclusief afgelegd op buitenlandse binnenwateren)	Numeriek
afstand totaal	afstand tussen plaats lading tot plaats lossing van een zending (in km), inclusief afgelegd op buitenlandse binnenwateren)	Numeriek
land lading	land waar goed geladen is	EU codering: 000 tm 999
land lossing	land waar goed gelost is	EU codering: 000 tm 999
landsdeel lading	landsdeel (Nederland) waar goed geladen is	1 t/m 7
landsdeel lossing	landsdeel (Nederland) waar goed gelost is	1 t/m 7
provincie lading	provincie waar (Nederland) waar goed geladen is	1 t/m 12
provincie lossing	provincie (Nederland) waar goed gelost is	1 t/m 12
COROP lading	COROP (Nederland) waar goed geladen is	010 t/m 999, zie tabel
COROP lossing	COROP (Nederland) waar goed gelost is	010 t/m 999, zie tabel
verkeersgebied lading	Verkeersgebied (Nederland) waar goed geladen is	01 t/m 54, zie tabel
verkeersgebied lossing	Verkeersgebied (Nederland) waar goed gelost is	01 t/m 54, zie tabel
ladingtonkilometer Nederland	Vervoerd gewicht (in tonnen) maal afstand Nederland (in km)	Numeriek
ladingtonkilometer totaal	Vervoerd gewicht (in tonnen) maal afstand totaal	Numeriek
Variabelen bij containergroep	***Definitie***	***Domein***
Grootte	Lengte van container (in voets)	20, 30, 40, 50,90
Aantal	Aantal containers in de groep	Numeriek
Beladen	Beladingstoestand van een container	0=leeg 1= beladen 2=deelladingen
Emballagegewícht	eigengewicht van een container	2,5 , 3 , 3,5
nummer?	Unieke identificatie van individuele container	??
Variabelen bij Telpunt	***Definitie***	***Domein***
Nummer	aanduiding van plaats waar telpunt ligt	8 cijferig nummer
Naam	naam van telpunt (veelal sluis)	Nvt
Vaarrichting	(wind)richting waarin het schip het telpunt passeert	N,O,Z,W
datum passeren	datum waarop een vaartuig een telpunt passseert en daar geregistreerd wordt	Datum
tijdstip passeren	tijdsaanduiding waarop een vaartuig een telpunt passseert en daar geregistreerd wordt	Tijd
Variabelen bij Reis	***Definitie***	***Domein***
reis-id	uniek identificatienummer van reis binnen IVS (dummy waarde voor andere bronnen)	Nvt

plaats lading	plek waar eerste goederen aan boord zijn gekomen	UNLO-codering
plaats lossing	plek waar laatste goederen van boord zijn gegaan	UNLO-codering
Datum	passagedatum bij geselecteerd telpunt	Datum
totale belading	som van vervoerd gewicht van alle zendingen op deze reis (in 1000 kg)	numeriek, 2 decimalen
afstand Nederland	afstand van plaats lading tot plaats lossing van een reis (in km), exclusief afgelegd op buitenlandse binnenwateren	Numeriek
afstand totaal	afstand van plaats lading tot plaats lossing van een reis (in km), exclusief afgelegd op buitenlandse binnenwateren	Numeriek
Oorsprongscode	indicatie van bronbestand record	IVS, Maandstaten, Zaandam
kantoor inklaring	code voor grensovergang waar schip Nederland binnenkomt	11 t/ 99, zie tabel
kantoor uitklaring	code voor grensovergang waar schip Nederland verlaat	11 t/ 99, zie tabel
Status	indicatie van de reeds ondergane bewerkingen van dit record in het verwerkingsproces (procesvariabele)	Nieuw, Fouten geconstateerd, Fout handmatig gecorrigeerd, Logich verwijderd, Verwerkt
Variabelen bij Schip	***Definitie***	***Domein***
Scheepsnummer	eenduidig nummer, gekoppeld aan één schip	8 cijferig
Naam	naamsaanduidigng van een schip	Nvt
Laadvermogen	maximal gewicht aan goederen (in 1000 kg) dat een voertuig mag vervoeren	Numeriek
type CRB	scheepstypecodering (Centrale Registratie Binnenschepen)	000 t/m 999, zie tabel
type RWS	scheepstypecodering (Rijkswaterstaat)	00 t/m 99, zie tabel
Lengte	lengte van een schip, van boeg tot achtersteven (in meters)	Numeriek
Breedte	breedte van een schip van bakboord naar stuurboord op het breedste punt (in cm)	Numeriek
Hoogte	hoogte van een schip van kiel tot hoogste punt (in cm)	Numeriek
Diepgang	maximale afstand van waterlijn tot kiel bij maximale belading (in cm)	Numeriek
Nationaliteit	codering van vlag waaronder het schip vaart	3 cijferige EU codering (zie tabel)
Laadvermogenklasse	maximal gewicht aan goederen (in 1000 kg) dat een voertuig mag vervoeren	0 t/m 9, zie tabel
Beladingscode	Beladingstoestand van een schip	1=leeg, 3=leeg,niet ontgast, 5=leeg,ontgast, 7=geladen, *enkelvoudige zending, 8= geladen, deelladingen, 9= geladen, containers*
laadvermogentonkilometer Nederland	Laadvermogen schip (in 1000 kg) maal afstand Nederland (in km)	Numeriek
Laadvermogentonkilometer totaal	Laadvermogen schip (in 1000 kg) maal afstand totaal (in km)	Numeriek