

HANDLEIDING STATISTIEKPRODUCTIE

11.07.2025



Vlaanderen
is sterk in cijfers

Inleiding	3
1 Wat is een statistiek?	3
1.1 Statistiek als een cijfer	3
1.1.1 Conceptuele definitie	4
1.1.2 Operationele definitie	6
1.1.3 Gebruikersbehoeften & kwaliteit	9
1.1.4 Attributen van cijfers	11
1.1.5 Samengevat	12
1.2 Statistiek als een cijferreeks	12
1.2.1 Wat is een statistiekreeks?	12
1.2.2 Concepten, dimensies en attributen	13
1.2.3 Conceptuele en operationele definitie	15
1.2.4 Samengevat	16
1.3 Statistiek als een tabelreeks	17
1.4 Experimentele statistieken	19
2 Documentatie — Metadata	20
3 Het statistiekproductieproces	20
3.1 Alles is data	20
3.2 Officiële Statistieken	20
3.2.1 Conceptenlijst	20
3.2.2 Hoe bepalen we concepten en dimensies?	21
3.2.3 Versies van statistiekreeksen	24
4 Ontsluiting statistieken	25
4.1 Statistiekreeksen	25
4.2 Cijferpagina's	26

INLEIDING

Één van de hoofdtaken van het netwerk Statistiek Vlaanderen en de Vlaamse Statistische Autoriteit is de productie en publicatie van openbare statistieken, ook wel Vlaamse Openbare Statistieken (VOS'en) genoemd. Maar wat verstaan we eigenlijk onder een statistiek en hoe produceren we die dan concreet? Om het netwerk goed te laten functioneren, is het cruciaal om met een gedeelde woordenschat te werken. Momenteel bestaan er echter grote verschillen in hoe termen en concepten binnen het netwerk worden gebruikt. Deze nota heeft daarom als doel enkele basisconcepten in de Vlaamse statistiekproductie te verduidelijken.

In deze handleiding wordt terminologie uit de GSIM-, GSBPM- en de SDMX-standaarden gebruikt, al verwijzen we voorlopig slechts sporadisch in enkele kanttekeningen naar deze standaarden. De integratie van beide standaarden kan in een volgende versie opgenomen worden.

1 WAT IS EEN STATISTIEK?

Voor we kunnen bespreken hoe we statistieken produceren, moeten we eerst afspreken wat we precies verstaan onder een statistiek. Wat beschouwen we als een statistiek en wat niet? Wat hebben we allemaal nodig om te spreken van een statistiek? Welke informatie is essentieel om statistiek te produceren? Hoe bepalen we de kwaliteit van een statistiek? Het antwoord op al deze vragen komt aan bod in deze paragraaf.

1.1 STATISTIEK ALS EEN CIJFER

Als je mensen vraagt wat ze verstaan onder een statistiek, zullen velen een statistiek op de eerste plaats definiëren als een cijfer. Zo'n definitie is echter onvolledig en onwerkbaar. Een statistiek houdt meer in dan enkel een cijfer. Statistiek gaat namelijk niet alleen over het berekenen van cijfers, het draait ook om de interpretatie ervan. Die interpretatie wordt gegeven door de definitie van de statistiek en in deze paragraaf zullen we zien dat er twee soorten definities bestaan: de conceptuele en operationele definitie. We zullen ook zien dat deze definities de kwaliteit van de statistiek bepalen en hoe dat gebeurt. Ten slotte, laat deze discussie ook toe om te af te bakenen wat experimentele statistieken zijn.

1.1.1 Conceptuele definitie

Een statistiek verwijst dus steeds naar een cijfer, maar een cijfer alleen bepaalt nog geen statistiek. Neem bijvoorbeeld het cijfer “525 935”. Dit cijfer is een voorbeeld van een bestaande statistiek. Toch zegt dit cijfer op zichzelf zeer weinig. Waarvan zijn er 525 935? Wat stelt dit getal voor? Zonder aanvullende informatie blijft het cijfer betekenisloos. Een *statistiek* bestaat daarom altijd uit twee elementen: het cijfer zelf én de betekenis ervan. Beide zijn even belangrijk en vormen één geheel. In dit geval betekent het cijfer 525 935: “het aantal inwoners van de gemeente Antwerpen op 1 januari 2019”. De betekenis van een statistiek noemen we de *conceptuele definitie*. Dit verwijst naar hoe een doorsnee gebruiker het cijfer interpreteert.

De conceptuele definitie vloeit steeds voort uit specifieke *gebruikersbehoeften*. Als we het aantal inwoners in Antwerpen publiceren als statistiek, doen we dit omdat bepaalde beleidsmakers, onderzoekers of andere burgers deze goed kunnen gebruiken. Zo verwijst een bepaald Vlaams decreet misschien naar het aantal inwoners in de Vlaamse gemeenten en zijn we als Statistiek Vlaanderen verplicht het aantal inwoners in Antwerpen te publiceren zodat de Vlaamse regering hierop beleid kan uitstippelen. Antwerpse beleidsmakers weten daarnaast waarschijnlijk ook graag hoeveel inwoners hun gemeente telt zodat ze hun diensten hierop kunnen afstemmen. Demografen gebruiken dit cijfer, samen met de bevolkingsgroottes in andere jaren en gemeenten, dan weer om bevolkingsgroei en -spreiding in Vlaanderen te onderzoeken.



Statistieken hoeven niet per definitie cijfers te zijn, het kunnen ook kwalitatieve observaties zijn of zelfs complexe informatie-objecten. Zo publiceert Agentschap Binnenlands Bestuur bijvoorbeeld per gemeente een statistiek over de top 10 andere gemeenten waarnaar kinderen heen pendelen voor school. Voor de stad Antwerpen kan de observatie op deze statistiek bijvoorbeeld de volgende tekstuele vector zijn:

‘[Antwerpen, Brasschaat, Lier, Brussel, Beveren, Mechelen, ...].’

Om het overzicht te bewaren blijven we hieronder echter verwijzen naar cijfers. Cijfers zijn immers veruit de meest voorkomende vorm van statistieken.

Merk op dat de conceptuele definitie verschillende componenten bevat, namelijk de grootte die we meten, dat is bevolkingsgrootte, de gemeenten waarin we dit meten, dat is Antwerpen, en het jaar waarin we dit meten, dat is 2019. Deze componenten noemen we *concepten*. Een conceptuele definitie bevat dus meestal meerdere concepten die we willen meten of gebruiken.

De afbakening van concepten in een conceptuele definitie is niet altijd duidelijk en is voor een stuk arbitrair. Zo kan je bijvoorbeeld cijfers publiceren over het aantal inwoners in Antwerpen per geslacht én het aantal inwoners per leeftijdsgroep maar zonder geslacht en leeftijd te kruisen (dat is dus, het aantal mannen, het aantal vrouwen, het aantal 0- tot 18-jarigen,...). Zijn geslacht en leeftijd in deze conceptuele definities twee verschillende concepten? Of

spreken we van één concept “bevolkingsgroep” (=“mannen”, “vrouwen”, “0- tot 18-jarigen”, “hoogopgeleiden”, “personen met buitenlandse herkomst”, ...). Als een ander voorbeeld, als we het aantal mannen en vrouwen publiceren in een gemeente. Hebben we hier twee concepten namelijk “grootheid” (=aantal) en “geslacht” (= “man” of “vrouw”) of slechts één concept “parameter” (=“aantal mannen” of “aantal vrouwen”)?



Er bestaan verschillende soorten concepten. Op de eerste plaats kunnen concepten concreet of abstract zijn.

- Een **concreet concept** verwijst naar een fenomeen dat direct observeerbaar of eenvoudig meetbaar is, vaak zonder complexe interpretatie of afleiding. Een voorbeeld van een concreet concept is het aantal huishoudens in een gemeente. Dit is een concreet concept: het gaat om een teller dat op basis van administratieve gegevens (bv. bevolkingsregister) rechtstreeks kan worden vastgesteld. Er is weinig discussie mogelijk over wat een huishouden is, zeker als er een duidelijke juridische of administratieve definitie wordt gehanteerd.
- Een **abstract concept** verwijst naar een idee of fenomeen dat niet rechtstreeks waarneembaar is. Het heeft een zekere mate van theoretische of interpretatieve lading en vereist meestal operationalisatie om meetbaar te worden.

Voorbeeld: Sociaal isolement Sociaal isolement is een abstract concept dat verwijst naar een toestand waarin iemand weinig of geen sociale contacten heeft of zich sociaal buitengesloten voelt. Dit kan je niet rechtstreeks meten; je moet het operationaliseren, bijvoorbeeld via een vragenlijst over het aantal sociale contacten, subjectieve gevoelens van eenzaamheid, deelname aan sociale activiteiten, enz.

Op de tweede plaats kunnen concepten enkelvoudig of meervoudig zijn.

- **Enkelvoudig concept** Een enkelvoudig concept verwijst naar een idee of kenmerk dat op één dimensie of aspect betrekking heeft. Het is relatief eenvoudig te definiëren en te meten, en het omvat één afgebakend betekenisveld.

Voorbeeld: Leeftijd Leeftijd is een enkelvoudig concept: het verwijst enkel naar het aantal jaren sinds iemands geboorte. Het is eenduidig te meten en wordt in de praktijk meestal als een numerieke variabele weergegeven (in jaren, maanden, ...).

- **Meervoudig concept** Een meervoudig concept verwijst naar een idee dat uit meerdere dimensies of onderliggende componenten bestaat. Je hebt meerdere indicatoren nodig om het concept volledig te vatten.

Voorbeeld: Welzijn Welzijn is een meervoudig concept, want het kan fysieke gezondheid, mentale gezondheid, sociale relaties, economische situatie, enz. omvatten. Om het te meten zijn verschillende indicatoren of meetinstrumenten nodig — bijvoorbeeld via samengestelde indexen of uitgebreide vragenlijsten.

De verschillende soorten concepten kunnen ook samen voorkomen zoals blijkt uit volgende voorbeelden:

- **Concreet & Enkelvoudig Voorbeeld: Leeftijd in jaren**

Uitleg: Dit is een direct observeerbaar, meetbaar kenmerk (concreet) en het gaat slechts over één dimensie (enkelvoudig). De meting is objectief en eenduidig.

- *Concreet & Meervoudig Voorbeeld: Woningkenmerken (bv. oppervlakte, aantal kamers, type dak)*

Uitleg: Deze kenmerken zijn stuk voor stuk observeerbaar (concreet), maar samen beschrijven ze een meervoudig concept zoals de "kwaliteit van een woning". Je hebt meerdere concrete indicatoren nodig om het volledig te vatten.

- *Abstract & Enkelvoudig Voorbeeld: Arbeidstevredenheid*

Uitleg: Dit is een subjectief, niet rechtstreeks observeerbaar gevoel (abstract), maar het wordt soms in onderzoek benaderd als één dimensie, bijvoorbeeld via een schaalvraag: "Hoe tevreden bent u over uw werk?". Dan behandel je het als een enkelvoudig concept, al kan je het ook complexer opvatten (zie hieronder).

- *Abstract & Meervoudig Voorbeeld: Sociaal isolement*

Uitleg: Dit is een abstract concept dat meerdere onderliggende aspecten omvat, zoals subjectieve eenzaamheid, frequentie van sociaal contact, deelname aan sociale activiteiten, enzovoort. Het vereist een set van vragen of indicatoren om het volledig te meten.

1.1.2 Operationele definitie

De conceptuele definitie beperkt zich doorgaans tot de betekenis die een doorsnee statistiekgebruiker aan het cijfer geeft. De conceptuele definitie is daarom meestal niet concreet over wat het cijfer precies weergeeft. In de zin "de Antwerpse bevolkingsgrootte neemt toe" is bijvoorbeeld "bevolkingsgrootte" enkel een abstract idee. Het woord vertelt immers niet wie je nu precies meetelt en wie niet om de grootte van de bevolking te bepalen. De conceptuele definitie van een cijfer vertelt dus weinig over hoe het cijfer precies werd berekend en exact geïnterpreteerd kan worden.

Naast de conceptuele definitie heeft elke statistiek daarom ook een *operationele definitie*. Deze operationele definitie beschrijft tot in detail hoe het cijfer exact is gemeten en geeft zo de precieze betekenis ervan weer. De operationele definitie is daarom meestal veel uitgebreider dan de conceptuele definitie.

Neem opnieuw het cijfer 525 935 om het aantal inwoners van Antwerpen in 2019 weer te geven. De volledige operationele definitie van van dit cijfer luidt:

"De grootte van de wettelijke bevolking op 1 januari 2019, 0.00 uur, van de gemeente Antwerpen (NIS 11002 in 2019). De wettelijke bevolking telt alle inschrijvingen in het bevolkingsregister en het vreemdelingenregister. Het bevolkingsregister bevat alle Belgen en buitenlanders die gemachtigd zijn tot vestiging op het Belgisch grondgebied. Het vreemdelingenregister bevat alle

buitenlanders die toegelaten of gemachtigd zijn tot een verblijf van meer dan 3 maanden op het Belgisch grondgebied, hetzij voor bepaalde of onbepaalde duur. Bepaalde categorieën buitenlanders (vb. diplomatiek en consulaire personeel) zijn vrijgesteld van inschrijving in de bevolkingsregisters. In sommige gevallen kunnen zij op eigen vraag wel ingeschreven worden. Enkel in dat geval worden zij meegerekend in de bevolkingscijfers.

Deze bevolkingsgrootte wordt gedefinieerd en aangeleverd door Statbel op basis van het Rijksregister van de natuurlijke personen, waar het bevolkingsregister en het vreemdelingenregister deel van uitmaken.

Het Rijksregister omvat verder ook een wachtregister voor asielzoekers en een wachtregister voor EU-burgers. Het wachtregister voor asielzoekers bevat alle verzoekers om internationale bescherming die worden ingeschreven door de Dienst Vreemdelingenzaken (DVZ). In 1995 besliste Statbel de personen in dit wachtregister niet meer mee te tellen bij de wettelijke bevolking. Pas nadat asielzoekers worden overgeschreven van het wachtregister naar het bevolkingsregister of het vreemdelingenregister, worden zij opgenomen in de bevolkingsstatistieken van Statbel. Zo'n overschrijving naar het bevolkingsregister of het vreemdelingenregister gebeurt na erkenning als vluchteling, na toekenning van een statuut subsidiaire bescherming, of na verwerving van een verblijfsvergunning om een andere reden.

Verder bevat het Rijksregister ook een wachtregister voor EU-burgers in afwachting van woonstcontrole. Deze personen worden evenmin meegeteld bij de wettelijke bevolking. Pas na woonstcontrole worden deze personen overgeschreven naar het vreemdelingenregister en worden zij meegeteld in de wettelijke bevolking."

TOEVOEGEN: elk concept wordt geoperationaliseerd

Zoals je ziet, de operationele definitie is alle een hele mond vol, zelfs voor een redelijk eenvoudige statistiek als de grootte van een bevolking. De lengte van een operationele definitie kan dan ook sterk variëren. Voor een eenvoudig cijfer afgeleid uit officiële registers, zoals de bevolkingsgrootte in Vlaamse gemeenten, kunnen enkele korte paragrafen volstaan. Bij complexere statistieken, zoals bijvoorbeeld de gemiddelde opinie van inwoners gemeten via een bevraging, is daarentegen een uitgebreid rapport nodig met alle details over, onder andere, het enquêtedesign, de steekproeftrekking, het ontwerp van de vragenlijst, de opvolging van respondenten, statistische correcties voor meet- en selectiefouten en de statistische analysemethoden.

Het cijfer 525 935 verwijst dus naar de bevolking op 1 januari 2019. Een vergelijkbaar maar ander cijfer kon worden berekend voor elke andere dag in het jaar of als een jaargemiddelde. Bovendien verwijst 525 935 naar de wettelijke bevolking gerapporteerd door Statbel en in die wettelijke bevolking worden personen uit de wachtregisters niet meegeteld. Een vergelijkbaar maar ander cijfer had ook kunnen worden berekend voor de verblijvende bevolking van Antwerpen waarin personen uit de wachtregisters wel worden meegeteld. Verder verwijst het cijfer 525 935 naar de gemeente Antwerpen zoals gedefinieerd door NIS-code 11002 in 2019. Door fusies, splitsingen of grensaanpassingen verandert het grondgebied van sommige gemeenten over de tijd; bijvoorbeeld, in 2025 fuseerde Antwerpen met Borsbeek, waardoor het grondgebied veranderde. We zouden opnieuw een vergelijkbaar maar ander cijfer kunnen berekenen voor alle personen in 2019 die op dat moment woonden in wat we nu als de gemeente Antwerpen beschouwen, inclusief Borsbeek dus.

Dit voorbeeld illustreert dat zelfs relatief eenvoudige conceptuele definities vaak vertaald kunnen worden naar meerdere operationele definities, elk met eigen operationele keuzes. Die keuzes kunnen leiden tot verschillende cijfers. Eurostat, bijvoorbeeld, rapporteert niet de wettelijke bevolking maar de gewoon verblijvende bevolking en publiceert daarom andere cijfers dan Stabel en Statistiek Vlaanderen over bevolkingsgroottes, ook al interpreteren de meeste gebruikers zowel de Europese als de Belgische en Vlaamse cijfers gewoon als de “bevolkingsgrootte”.

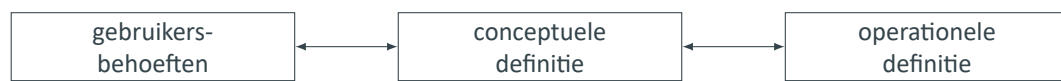
Het eerste doel van de operationele definitie is dus duidelijkheid te scheppen over de gemaakte keuzes om een statistiek te berekenen zodat iedereen hetzelfde cijfer kan reproduceren. Als twee onderzoekers met dezelfde operationele definitie verschillende cijfers bekomen, is de definitie onvolledig of onnauwkeurig.

Het tweede doel van de operationele definitie is duidelijkheid te scheppen over de vergelijkbaarheid van cijfers. Je kan bijvoorbeeld niet zomaar de grootte van de wettelijke Antwerpse bevolking in 2019 vergelijken met de grootte van de verblijvende Antwerpse bevolking in 2020. Dit zou immers kunnen leiden tot onjuiste conclusies over bevolkingsgroei of -krimp, ook al worden beide cijfers conceptueel geïnterpreteerd als de Antwerpse bevolkingsgrootte.



In theorie leidt een operationele definitie steeds tot één exact resultaat. In de praktijk is dit echter niet steeds het geval. Als het productieproces bijvoorbeeld een willekeurige steekproef of simulatie bevat, kan het cijfer variëren. Toch blijft de operationele definitie als idee ook in zo’n situaties overeind. Twee onderzoekers zouden in zo’n situatie op basis van dezelfde operationele definitie gemiddeld steeds tot hetzelfde cijfer moeten komen wanneer zij hun productieproces blijven herhalen. Het gebied van de inferentiële statistiek biedt hiervoor het theoretisch kader, maar dat gaat voorbij het doel van deze nota.

Figuur 1: De kwaliteit van een statistiek wordt gedefinieerd door de verhouding tussen drie componenten.



Je kan je afvragen waarom we eigenlijk nog conceptuele definities opstellen als ze, in tegenstelling tot de operationele definitie, niet accuraat zijn? Toch zijn er goede redenen om voor elke statistiek een goede conceptuele definitie te bepalen. Ten eerste is de conceptuele definitie een pak handiger in gebruik dan de operationele definitie binnen een narratief of theorievorming. Statistiekgebruikers beperken zich daarom doorgaans tot de conceptuele definitie om cijfers te benoemen en te interpreteren, ook al zijn er binnen dezelfde conceptuele definitie meerdere operationele definities mogelijk. Iemand die praat over het aantal inwoners in Antwerpen in 2019 zal dit eenvoudigweg de “Antwerpse bevolkingsgrootte” noemen en niet de hele operationele definitie afratelen. Ten tweede hebben de nuances in de operationele definitie vaak weinig invloed op beleidskeuzes of onderzoeksconclusies die gebruikers maken op basis van de cijfers. Of je nu de feitelijke of de verblijvende bevolking gebruikt om de bevolkingsomvang in Antwerpen te meten, en of je dit nu doet aan het begin of in het midden van het kalenderjaar, de meeste gebruikers zullen min of meer dezelfde conclusies trekken over bevolkingsgroei als ze deze cijfers over de jaren heen vergelijken.

1.1.3 Gebruikersbehoeften & kwaliteit

We zagen dus dat een statistiek wordt gedefinieerd door 3 zaken:

- de gebruikersbehoeften
- de conceptuele definitie
- de operationele definitie

Wanneer er een wanverhouding ontstaat tussen deze componenten ontstaat er een kwaliteitsprobleem. Dit kan tot uiting komen in verschillende vormen:

- De conceptuele definitie is niet in lijn met de algemene gebruikersbehoeften.
- De conceptuele definitie is niet in lijn met gebruikersbehoeften van sommige gebruikers
- De operationele definitie is niet in lijn met de conceptuele definitie.

We bespreken deze situaties één voor één hieronder.

De conceptuele definitie is niet in lijn met de algemene gebruikersbehoeften In sommige situaties kan er een wanverhouding ontstaan tussen gebruikersbehoeften en de operationele definitie van een

statistiek. Zo kunnen gebruikers bijvoorbeeld het cijfer 525 935 systematisch interpreteren als de grootte van de verblijvende Antwerpse bevolking in plaats van de wettelijke bevolking omdat ze beleidsmatig ook voornamelijk de grootte van de verblijvende bevolking nodig hebben. In zo'n situatie spreken we van een *kwaliteitsprobleem* omdat de operationele definitie van onze cijfers niet in overeenstemming is met de gebruikersbehoeften.

Voor cijfers waarbij we zo'n kwaliteitsprobleem opmerken, moeten we onderzoeken of we onze operationele definitie kunnen aanpassen aan de behoeften. In bovenstaande situatie zou het bijvoorbeeld aangeraden zijn om als Statistiek Vlaanderen voortaan Eurostat te volgen en de grootte van de verblijvende bevolking te publiceren in plaats van de wettelijke bevolking.

De conceptuele definitie is niet in lijn met gebruikersbehoeften van sommige gebruikers In vele situaties is een evaluatie van gebruikersbehoeften en kwaliteit van een statistiek echter geen zwart-wit-verhaal. Er kunnen zich immers situaties voordoen waarin de voorgestelde operationele definitie voldoet voor de ene gebruiker maar niet voor een andere gebruiker. Zo kan onze statistiek over bevolkingsgrootte op basis van de wettelijke bevolking voldoende zijn voor heel wat gebruikers, behalve voor een specifieke beleidsmaker die het woonbeleid in Antwerpen moet bepalen. Deze beleidsmaker heeft misschien nood aan exacte cijfers over de verblijvende bevolking in plaats van de wettelijke bevolking. Ook in zo'n geval is er een gebruikersbehoefte waar we met de statistiek "Aantal inwoners in Antwerpen" niet aan voldoen, ook al voldoet deze statistiek wel voor andere gebruikersbehoeften.

In situaties met verschillende niet-overlappende gebruikersbehoeften kunnen we besluiten niet langer één statistiek te publiceren maar verschillende statistieken. Zo kunnen we besluiten niet één statistiek te publiceren over "het aantal inwoners in Antwerpen", maar wel twee afzonderlijke statistieken: "de wettelijke bevolkingsgrootte van Antwerpen" en "de verblijvende bevolkingsgrootte van Antwerpen". We voldoen dan aan de specifieke noden van alle gebruikers. Merk op dat we in zo'n situaties zowel de conceptuele definities als de bijhorende operationele definities verder verfijnen. Uiteraard moeten dit soort keuzes steeds gebeuren vanuit praktische en pragmatische overwegingen in functie van de middelen en het personeel die we ter beschikking hebben. We kunnen in de bovenstaande situatie bijvoorbeeld ook besluiten om niet te voldoen aan de noden van deze éne specifieke beleidsmaker omdat dit te veel energie zou vragen voor een te kleine return.

De operationele definitie is niet in lijn met de conceptuele definitie
AANVULLEN

Bovenstaande situaties tonen dat het essentieel is om voor elke statistiek niet alleen de conceptuele en operationele definitie, maar ook de gebruikersbehoeften grondig in kaart te brengen. Wie gebruikt de statistieken? Welk decreet dwingt ons ertoe deze statistieken te verzamelen? Voor welke beleidsbeslissingen of welk onderzoek zijn deze statistieken relevant? ...? Hoewel onderzoek naar deze vragen geen exacte wetenschap is, is het noodzakelijk om de kwaliteit van statistieken te kunnen beoordelen.

We moeten ook opmerken dat kwaliteit in de brede zin van het woord kan worden bekeken. Een operationele definitie kan inhoudelijk aansluiten bij een gebruikersbehoefte, maar als de berekening te lang duurt en de statistiek te laat beschikbaar is, verlaagt dit alsnog de kwaliteit. Het is dus belangrijk een evenwicht te vinden tussen verschillende elementen van kwaliteit zoals inhoudelijke nauwkeurigheid, gebruiksvriendelijkheid en tijdige levering. Een overzicht van alle kwaliteitseisen voor statistieken wordt gegeven in de *Praktijkcode voor Europese Statistieken*¹.



Gebruikersbehoeften, conceptuele en operationele definitie komen weliswaar aan bod in informatiestandaarden zoals de SIMS en de GSBPM maar op een zeer onduidelijke en onoverzichtelijke manier. In de SIMS worden de gebruikersbehoeften pas opgelijst in categorie 12 terwijl dit de start vormt van een statistiekproductieproces. De operationele definitie zit in de SIMS dan weer op een zeer onsamenhangende manier verspreid over verschillende categorieën terwijl de conceptuele definitie volledig ontbreekt. De GSBPM doet het op dit vlak beter. Het beschrijft een proces dat vertrekt vanuit gebruikersbehoeften en eindigt met een evaluatie. Het expliciteert echter onvoldoende dat de evaluatie gestoeld moet zijn op de gebruikersbehoeften en alle operationele keuzes die onderweg werden gemaakt.

1.1.4 Attributen van cijfers

Naast de operationale definitie van een cijfer, kan een cijfer gekenmerkt worden door andere informatie, die geen betekenis geeft aan het cijfer. Dit worden *attributen* genoemd.

Een belangrijke informatiepunt is de vertrouwelijkheid van het cijfer. Kan het cijfer worden gedeeld met het breed publiek, met collega's of is het vertrouwelijk en moet het veilig bewaard worden.

Een ander attribuut is de eigenaar van het cijfer. De eigenaar van een cijfer verwijst naar wie verantwoordelijk is voor dat cijfer. Deze kan

¹ zie ec.europa.eu/eurostat/documents/4031688/9394211/KS-02-18-142-NL-N.pdf

Tabel 2: Een statistiek bestaat uit een cijfer, een conceptuele definitie die voortvloeit uit gebruikersbehoeften, en een operationele definitie die vertelt hoe het cijfer exact wordt gemeten. Een wanverhouding tussen gebruikersbehoeften en de operationele definitie bepaalt de kwaliteit van de statistiek.

Gebruikersbehoeften:	→ Decreet XX.XX verwijst naar het aantal inwoners in Vlaamse gemeenten in de context van ... → Agentschap Binnenlands Bestuur publiceert cijfers over het aantal inwoners in gemeenten om informatie te geven over ... → Academische onderzoekers vragen cijfers over het aantal inwoners in gemeenten om onderzoek te voeren over ... → ...
Conceptuele definitie:	Aantal inwoners van de gemeente Antwerpen in 2019
Operationele definitie:	De grootte van de wettelijke bevolking op 1 januari 2019 0.00 uur in de gemeente Antwerpen (Statbel NIS-code 11002 in 2019). De data worden aangeleverd door Statbel op basis van het Rijkregister van de natuurlijke personen. De wettelijke bevolking verwijst naar ...
Cijfer:	525 935
Kwaliteitsevaluatie:	De gebruikers hebben ook nood aan de grootte van de verblijvende bevolking naast de wettelijke bevolking

1.1.5 Samengevat

Samengevat, voor elke statistiek die we produceren en publiceren hebben we nood aan:

1. de gebruikersbehoeften,
2. een conceptuele definitie op basis van de gebruikersbehoeften,
3. een operationele definitie,
4. het cijfer, en
5. een kwaliteitsevaluatie van de operationele definitie in functie van de gebruikersbehoeften.

Tabel 2 vat deze basisinformatie samen aan de hand van het Antwerps voorbeeld. Als de gebruikersbehoeften van een statistiek nog niet in kaart zijn gebracht en er daardoor ook geen kwaliteitsevaluatie kan worden uitgevoerd, is de statistiek een experimentele statistiek.

1.2 STATISTIEK ALS EEN CIJFERREEKS

1.2.1 Wat is een statistiekreeks?

In de vorige paragraaf werd uitgelegd hoe een statistiek verwijst naar één enkel cijfer met een bijbehorende conceptuele en operationele definitie ontstaan

Tabel 3: We publiceren doorgaans geen individuele statistieken maar statistiekreeksen.

Conceptuele definitie	Cijfer
Aantal inwoners in Aalst in 2019	86 445
Aantal inwoners in Aalter in 2019	28 906
Aantal inwoners in Aarschot in 2019	30 115
Aantal inwoners in Aartselaar in 2019	14 293
Aantal inwoners in Affligem in 2019	13 228
Aantal inwoners in Alken in 2019	11 499
Aantal inwoners in Alveringem in 2019	5 047
Aantal inwoners in Antwerpen in 2019	525 935
Aantal inwoners in Anzegem in 2019	14 716
...	...
Aantal inwoners in Zwijndrecht in 2019	19 056

vanuit specifieke gebruikersbehoeften. In de praktijk produceren we echter veel statistieken met sterk overlappende gebruikersbehoeften en definities. Zo publiceren we niet alleen het aantal inwoners voor de gemeente Antwerpen, maar ook voor alle andere Vlaamse gemeenten, de drie gewesten en heel België. Aparte definities opstellen en beheren voor al deze cijfers is niet efficiënt.

Om efficiëntie te verhogen werken we daarom met *statistiekreeksen* of cijferreeksen: verzamelingen van statistieken die grotendeels dezelfde conceptuele en operationele definities delen. Tabel 3 toont bijvoorbeeld een statistiekreeks met de inwonersaantallen van alle Vlaamse gemeenten.

1.2.2 Concepten, dimensies en attributen

Aangezien de conceptuele en operationele definities van individuele cijfers in statistiekreeksen overlappen, kunnen we deze definities opsplitsen in verschillende onderdelen. Deze verschillende onderdelen van de definities worden de *concepten* genoemd. Concepten zijn eigenlijk gewoon variabelen, kenmerken waar de statistieken naar refereren en die kunnen variëren tussen statistieken. Bij de statistiekreeks in tabel 3 kunnen we bijvoorbeeld drie concepten onderscheiden: (1) de parameter “aantal inwoners”, (2) het jaartal “2019”, en (3) de gemeente. Niet alle statistieken die we publiceren geven immers het aantal inwoners als parameter weer, en evenmin geven ze allemaal informatie voor het jaar 2019. De statistiekreeks in tabel 3 toont bovendien al aan hoe gemeente varieert over verschillende statistieken.

Binnen één statistiekreeks wordt een onderscheid gemaakt tussen twee soorten concepten. Sommige concepten helpen elk cijfer uniek te identificeren, terwijl anderen enkel bijkomende informatie geven over de cijfers. De concepten die de cijfers helpen identificeren binnen een statistiekreeks

bepalen de *dimensies* van de statistiekreeks. Ze worden ook wel *identificeerders* (identifiers) genoemd. In tabel 3 is er slechts één dimensie, namelijk de gemeente, aangezien elk cijfer het aantal inwoners van een andere gemeente weergeeft en de cijfers enkel van elkaar verschillen door de de gemeente waarnaar ze verwijzen.

De concepten die enkel bijkomende informatie geven zonder de cijfers verder te identificeren worden daarentegen *attributen* genoemd. Attributen worden daarom ook soms *beschrijvers* (describers) genoemd. In tabel 3 zijn de parameter en het jaartal de attributen, omdat elk cijfer verwijst naar een inwonersaantal in 2019 en niets anders. Attributen kunnen informatie bevatten voor de hele statistiekreeks, maar kunnen evengoed enkel informatie bevatten voor één enkel cijfer of een beperkte groep van cijfers binnen de statistiekreeks. Een voorbeeld van zo'n laatste situatie is informatie over de vertrouwelijkheid van cijfers. Één groep cijfers in een tabel kan beschouwd worden als vrij te publiceren terwijl een andere groep niet openbaar mag gemaakt worden omdat ze vertrouwelijke informatie kunnen ontsluiten. De aanduiding van deze vertrouwelijkheid identificeert de cijfers op zich niet maar varieert wel van cijfer tot cijfer.

Onthoud echter dat attributen dimensies kunnen worden wanneer verschillende statistiekreeksen worden samengevoegd. Om bijvoorbeeld bevolkingsdichtheden te berekenen, moeten de cijfers uit tabel 3 worden gecombineerd met cijfers over de oppervlakte van alle gemeenten. In die samengestelde tabel wordt de parameter een dimensie in plaats van een attribuut, omdat sommige cijfers verwijzen naar de parameter 'aantal inwoners', terwijl anderen verwijzen naar de parameter "oppervlakte" van de gemeenten. Hetzelfde gebeurt met het concept jaartal wanneer we aantallen inwoners combineren over verschillende jaren heen voor een tijdreeks. Omgekeerd kan een dimensie ook een attribuut worden wanneer we een tabel opsplitsen volgens die dimensie. Bijvoorbeeld, als we longitudinale data op jaarbasis opsplitsen in aparte tabellen per jaartal, wordt jaartal in die nieuwe tabellen slechts een attribuut in plaats van een dimensie. Het verschil tussen attributen en dimensies is dus relatief want het hangt af van welke tabel je precies bekijkt. Over al mogelijke statistieken bekeken zijn alle concepten dimensies want geen enkele operationele definitie is dezelfde.



We hebben hier een hiaat in het SDMX model blootgelegd. De documentatie van SDMX bespreekt nergens hoe attributen dimensies kunnen worden of vice versa wanneer datasets met statistiekreeksen worden samengevoegd of uitgesplitst. Om deze reden is de uitwerking van attributen in het model ook veel beperkter dan de uitwerking van dimensies. Hierdoor kan SDMX moeilijker gelinkt worden aan de ideeën rond conceptuele definities, operationele definities en kwaliteit van statistiekreeksen.

Sterker nog, binnen SDMX wordt parameter nooit duidelijk beschouwd als een dimensie waardoor je verschillende parameters zoals bevolkingsgrootte en

oppervlakte moeilijk in één tabel kan combineren zonder in problemen te komen over de definitie van dimensies en attributen.

1.2.3 Conceptuele en operationele definitie

Net zoals elke afzonderlijke statistiek worden statistiekreeksen ook beschreven via een conceptuele en een operationele definitie. Dit vergemakkelijkt informatiebeheer over de statistiekreeksen.

De *conceptuele definitie* verwijst naar de interpretatie die een doorsnee gebruiker aan de statistiekreeks toekent. De statistiekreeks in tabel 3 kan bijvoorbeeld conceptueel worden gedefinieerd als het “aantal inwoners in gemeenten in het Vlaamse Gewest in 2019”. Deze definitie sluit aan bij specifieke gebruikersbehoeften, zoals bijvoorbeeld beschreven in tabel 2.

De *operationele definitie* van een statistiekreeks beschrijft hoe de cijferreeks precies werd verzameld en gereproduceerd kan worden. Bovendien specificeert de operationele definitie steeds welke waarden de attributen en dimensies binnen de reeks precies aannemen. Voor tabel 3 luidt de operationele definitie bijvoorbeeld: “De grootte van de wettelijke bevolking op 1 januari 2019 per gemeente volgens de NIS-code-indeling in 2019. De wettelijke bevolking omvat...” Deze definitie benoemt heel duidelijk ‘gemeente’ als dimensie en geeft aan hoeveel gemeenten de reeks omvat, namelijk de 300 Vlaamse gemeenten volgens de NIS-code-indeling van 2019, en bijvoorbeeld niet de 285 Vlaamse gemeenten vanaf 2025. Daarnaast beschrijft de definitie ook heel duidelijk de waarde van de attributen. De reeks bevat namelijk cijfers over de grootte van de wettelijke bevolking en bijvoorbeeld niet de verblijvende bevolking, en voor het jaar 2019 en geen ander kalenderjaar.

Door de exacte beschrijving van dimensies bepaalt de operationele definitie bovendien strikt uit hoeveel cijfers een statistiekreeks bestaat, zelfs als sommige cijfers een missende of versluisde waarde hebben. De reeks in tabel 3 bestaat bijvoorbeeld uit 300 cijfers, voor elke Vlaamse gemeente één. Als er in de reeks ook totaalcijfers zouden worden gepubliceerd voor Vlaanderen, Wallonië, Brussel en België, verandert de dimensie ‘gemeente’ naar het ruimer concept ‘geografisch gebied’ met 304 mogelijke waarden en dikt de reeks tot evenveel cijfers aan. Splitsen we de cijfers verder uit naar mannen en vrouwen, dan wordt de dimensie ‘geslacht’ toegevoegd met twee mogelijke waarden (man en vrouw) en bevat de reeks plots 608 cijfers. Inclusief de totalen over beide geslachten komt het totaal op 912 cijfers. Voeg daar nog het percentage mannen en vrouwen per gemeente aan toe, dan bevat de reeks 1520 cijfers.

Een kleine tip om de operationele definitie duidelijk uit te schrijven: Maak slim gebruik van voorzetsels zoals “voor”, “in” of “per” om de attributen en dimensies aan te duiden. Zo kan je bijvoorbeeld een statistiekreeks definiëren

als het “aantal inwoners IN het jaar 2019 PER geografisch gebied (Vlaamse gemeenten, Belgische gewesten en heel België volgens de NIS-code-indeling van 2019), VOOR elk geslacht (mannen, vrouwen, totaal), en PER leeftijdsgroep (0-18, 19-30, 31-65, 66+ jaar, totaal).”

Hoewel het bij de conceptuele definitie minder strikt is om alle attributen en dimensies expliciet te vermelden, is dit ook een goede praktijk. Voor de statistiekreeks in tabel 3 kunnen we bijvoorbeeld kiezen tussen de conceptuele definities “bevolkingsomvang” of “aantal inwoners in de Vlaamse gemeenten in 2019”. De tweede definitie is uiteraard een pak duidelijker als omschrijving. De eerste definitie is anderzijds een pak beknopter en gemakkelijker te citeren maar wordt uiteraard dubbelzinnig als we ook tabellen publiceren voor andere kalenderjaren (tenzij we al deze tabellen samenvoegen natuurlijk, maar dan zitten we weer in een ander verhaal).

Net zoals bij individuele statistieken bepaalt een wanverhouding tussen gebruikersbehoeften en de operationele definitie mee de kwaliteit van een statistiekreeks. Als de operationele definitie niet in lijn ligt met de behoeften van gebruikers moet de reeks worden aangepast of worden geschrapt.

De presentatie van een statistiekreeks gebeurt doorgaans niet zoals in tabel 3. Attributen worden vaak uit de rijen verwijderd en opgenomen in de titel of conceptuele definitie van de statistiekreeks. Tabel 4 toont een meer gangbare weergave van een uitgebreidere cijferreeks. In deze tabel zie je dat de concepten “aantal inwoners” en jaartal “2019” enkel worden vermeld in de titel van de tabel. De tabel telt verder drie dimensies. Dimensies gemeente en geslacht staan in de rijen. De derde dimensie is de statistische parameter die een verschil maakt tussen aantallen en percentages. Deze dimensie staat in de kolommen.



Binnen de huidige SDMX-standaard stellen we statistiekreeksen voor met slechts één kolom voor de cijfers. De structuur in tabel 3 zou hiervoor gepivoteerd moeten worden. Hoogstwaarschijnlijk versoepelt de SDMX-standaard echter in de toekomst waardoor verschillende kolommen cijfers kunnen bevatten.

1.2.4 Samengevat

Samengevat, een statistiekreeks is een reeks van statistieken waarbij gebruikersbehoeften, conceptuele en operationele definities overlappen. Elke statistiekreeks is nauwkeurig gedefinieerd via haar operationele definitie. Deze operationele definitie vertelt exact hoeveel cijfers de statistiekreeks bevat, wat deze cijfers betekenen en hoe de cijfers werden bekomen. Een wanverhouding tussen de operationele definitie en de gebruikersbehoeften is een van de kwaliteitsindicatoren van de statistiekreeks volgens de praktijkcode.

Tabel 4: De conceptuele definitie van een statistiekreeks verwijst naar de interpretatie van een doorsnee gebruiker terwijl de operationele definitie alle concepten (attributen en dimensies) van de reeks duidelijk beschrijft.

AANTAL MANNEN EN VROUWEN IN VLAAMSE GEMEENTEN IN 2019.			
Het aantal en het percentage mannen en vrouwen in de wettelijke bevolking op 1 januari 2019 per gemeente volgens de NIS-code-indeling in 2019.			
Gemeente	Geslacht	Aantal	Percentage
Aartselaar	man	7089	49.6
Aartselaar	vrouw	7204	50.4
Antwerpen	man	262921	50.0
Antwerpen	vrouw	263014	50.0
Boechout	man	6506	49.0
Boechout	vrouw	6760	51.0
Boom	man	9024	49.5
Boom	vrouw	9220	50.5
Borsbeek	man	5270	48.6
Borsbeek	vrouw	5584	51.4

Met Statistiek Vlaanderen leggen we op voorhand vast welke statistiekreeksen we publiceren. Deze statistiekreeksen zijn terug te vinden in het Vlaams Statistisch Programma (VSP-lijst). Voor elke statistiek in de VSP-lijst voorzien we ten minste volgende informatie:

1. de conceptuele definitie (die gebruikt wordt als handig label voor de statistiekreeks),
2. het versienummer
3. de operationele definitie inclusief een beschrijving van alle concepten (attributen en dimensies) inclusief de exacte waarden van die concepten,
4. een overzicht van de gebruikersbehoeften, en
5. een kwaliteitsevaluatie van de operationele definitie in functie van de gebruikersbehoeften.

Tabel 5 toont hoe de VSP-lijst er minimaal zou moeten uitzien. In de praktijk zullen we uiteraard informatie over statistiekreeksen niet bewaren in één gigantische tabel maar werken met informatiefiches per statistiekreeks gebaseerd op de single integrated metadata structure (SIMS)².

1.3 STATISTIEK ALS EEN TABELREEKS

Voorbeeld: bevolking naar leeftijd en geslacht — bevolkingspyramide Hieruit kunnen heel wat andere cijferreeksen worden berekend - aantal inwoners

² ec.europa.eu/eurostat/web/metadata/reference-metadata-reporting-standards

Tabel 5: Het Vlaams Statistisch Programma (VSP) definieert een lijst van statistiekreeksen die Statistiek Vlaanderen publiceert. Elke statistiekreeks in deze reeks beschikt over een duidelijke conceptuele en operationele definitie, opgelijste gebruikersbehoeften en een kwaliteitsevaluatie.

<p>Aantal mannen en vrouwen, versie 1</p> <p><i>gebruikersvoorwaarden:</i></p> <ul style="list-style-type: none"> → Decreet XX.XX verwijst naar het aantal inwoners in Vlaamse gemeenten om beleid te voeren over ... → Agentschap Binnenlands Bestuur publiceert cijfers over het aantal inwoners in gemeenten als ... → Academische onderzoekers vragen cijfers over het aantal inwoners in gemeenten om onderzoek te voeren over ... → ... <p><i>Operationele definitie:</i> Het aantal en het percentage mannen en vrouwen in de feitelijke bevolking (aantal geregistreerde inwoners in het Rijksregister inclusief personen in het wachtregister en ambassadeurs) op 1 januari van elk kalenderjaar vanaf 2005 per Vlaamse gemeente volgens de NIS-code-indeling in 2019 = 600 cijfers per jaar.</p> <p><i>Kwaliteit:</i> De gebruikers hebben eerder nood aan de grootte van de verblijvende bevolking in plaats van de feitelijke bevolking.</p> <p><i>Nota:</i> Stopgezet in 2022, wegens aanpassing definitie.</p>
<p>Aantal mannen en vrouwen, versie 2</p> <p><i>gebruikersvoorwaarden:</i></p> <ul style="list-style-type: none"> → Decreet XX.XX verwijst naar het aantal inwoners in Vlaamse gemeenten om beleid te voeren over ... → Agentschap Binnenlands Bestuur publiceert cijfers over het aantal inwoners in gemeenten als ... → Academische onderzoekers vragen cijfers over het aantal inwoners in gemeenten om onderzoek te voeren over ... → ... <p><i>Operationele definitie:</i> Het aantal en het percentage mannen en vrouwen in de verblijvende bevolking (aantal geregistreerde inwoners in het Rijksregister inclusief personen in het wachtregister en ambassadeurs en personen die minder dan drie maanden in België verblijven) op 1 januari van elk kalenderjaar vanaf 2005 per Vlaamse gemeente volgens de NIS-code-indeling in 2019 = 600 cijfers per jaar.</p> <p><i>Kwaliteit:</i> Onderzoek XX toont aan dat er geen problemen zijn met deze statistiekreeks.</p> <p><i>Nota:</i> Stopgezet in 2025, wegens aanpassing gemeenten door fusies.</p>
<p>Aantal mannen en vrouwen, versie 3</p> <p><i>gebruikersvoorwaarden:</i></p> <ul style="list-style-type: none"> → Decreet XX.XX verwijst naar het aantal inwoners in Vlaamse gemeenten om beleid te voeren over ... → Agentschap Binnenlands Bestuur publiceert cijfers over het aantal inwoners in gemeenten als ... → Academische onderzoekers vragen cijfers over het aantal inwoners in gemeenten om onderzoek te voeren over ... → ... <p><i>Operationele definitie:</i> Het aantal en het percentage mannen en vrouwen in de verblijvende bevolking (aantal geregistreerde inwoners in het Rijksregister inclusief personen in het wachtregister en ambassadeurs en personen die minder dan drie maanden in België verblijven) op 1 januari van elk kalenderjaar vanaf 2005 per Vlaamse gemeente volgens de NIS-code-indeling in 2025 = 570 cijfers per jaar.</p> <p><i>Kwaliteit:</i> Onderzoek XX toont aan dat er geen problemen zijn met deze statistiekreeks.</p>
<p>Tewerkstelling in hoogtechnologische sector, versie 1</p> <p><i>gebruikersvoorwaarden:</i></p> <ul style="list-style-type: none"> → Decreet XX.XX verwijst naar Tewerkstelling in hoogtechnologische sector in kader van ... → De cijferpagina met cijfers over tewerkstelling in hoogtechnologische sector wordt X aantal keer per jaar geraadpleegd. → ... <p><i>Operationele definitie:</i> Percentage van de hele werkende bevolking aan de slag in de hoogtechnologische sector op 1 januari van elk kalenderjaar vanaf 2005 per Vlaamse gemeente en voor het hele Vlaamse gewest volgens de NIS-code-indeling in 2019. De werkende bevolking omhelst ...De hoogtechnologische sector bestaat uit- bedrijven ...De data worden verzameld via de Enquête naar de Arbeidskrachten (EAK) door Statbel. In deze enquête wordt data verzameld door ...= 301 cijfers per jaar.</p> <p><i>Kwaliteit:</i> Onderzoek YY toont aan dat er geen problemen zijn met deze statistiekreeks.</p>
<p>Drinkwaterkwaliteit, versie 1</p> <p><i>gebruikersvoorwaarden:</i></p> <ul style="list-style-type: none"> → Decreet XX.XX verwijst naar drinkwaterkwaliteit in kader van ... → ... <p><i>Operationele definitie:</i> Conformiteitspercentage van het kraantjeswater in heel Vlaanderen per jaar. Het conformiteitspercentage wordt berekend door Vlaamse Milieumaatschappij (VMM) op basis van het totale aantal analyses en het totale aantal vastgestelde normoverschrijdingen voor volgende parameters: ...= 1 cijfer per jaar.</p> <p><i>Kwaliteit:</i> Volgens rapport ZZ ontstaan er kleine onzekerheidsfouten in de meting van parameter x, waardoor het werkelijke conformiteitspercentage kan afwijken met 0,2%. Deze fluctuatie heeft slechts beperkte invloed op de kwaliteit waardoor geen herdefiniëring nodig is.</p>
<p>....., versie ...</p> <p><i>gebruikersvoorwaarden:</i> ...</p> <p><i>Operationele definitie:</i> ...</p> <p><i>Kwaliteit:</i> ...</p> <p>.....</p>

naar geslacht - percentage volgens geslacht - aantal inwoners naar leeftijd - percentage per leeftijd - percentage 65+'ers - vergrijzing - ...

Opnieuw conceptuele en operationele definitie nodig

1.4 EXPERIMENTELE STATISTIEKEN

Sommige statistieken ontstaan niet vanuit geobserveerde gebruikersbehoeften, maar vanuit veronderstellingen hierover. Dit kan voortkomen uit de inhoudelijke interesse van een collega of door nieuwe dataverzamelings- of analysemethoden. Deze statistieken noemen we *experimentele statistieken*.

Experimentele statistieken verwijzen dus niet enkel naar statistieken ontwikkeld met de allernieuwste analysetechnieken die recent op de markt werden gebracht. Eurostat definieert experimentele statistieken als “Experimental statistics use new data sources and methods to better respond to our users’ needs in a timely manner.” Wat “new data sources and methods” betekent in deze definitie is echter rekbaar. Zo kan een nieuwe vraag in een bevraging al beschouwd worden als een nieuwe databron en kunnen we deze vraag gebruiken om een nieuwe experimentele statistiek te ontwikkelen waarover we enkel aannames kunnen maken over de gebruikersbehoeften.

Bij gebrek aan concrete gebruikersbehoeften kan de kwaliteit van experimentele statistieken uiteraard niet meteen worden geëvalueerd. Kwaliteit verwijst immers, zoals eerder al aangehaald, naar de overeenkomst tussen de gebruikersbehoeften en de operationele definitie van een statistiek.

Daarom is het noodzakelijk om voor elke experimentele statistiek een deadline voor evaluatieonderzoek rond gebruikersbehoeften af te spreken (bijvoorbeeld vijf jaar na de eerste publicatie). Dit evaluatieonderzoek beantwoordt vragen zoals:

- Wordt de statistiek gebruikt, en zo ja, waarvoor?
- Sluit het gebruik en de interpretatie aan bij de operationele definitie?
- Zijn er suggesties van gebruikers om de statistiek beter af te stemmen op hun behoeften?

Eens gebruikersbehoeften in kaart zijn gebracht, kan de kwaliteit van de statistiek verder worden geëvalueerd. Afhankelijk van de resultaten kan de statistiek nadien:

- worden geüpgraded tot een volwaardige officiële statistiek;
- worden aangepast op basis van de bevindingen; of
- worden stopgezet indien de kwaliteit onvoldoende is.



Het groot deel van de statistieken van de VSA zijn vermoedelijk experimenteel, aangezien gebruikersbehoeften niet systematisch zijn vastgelegd. Hierdoor is een kwaliteitsbeoordeling van deze statistieken momenteel moeilijk. Onderzoek naar gebruikersbehoeften dringt zich op bij deze statistieken.

2 DOCUMENTATIE — METADATA

3 HET STATISTIEKPRODUCTIEPROCES

3.1 ALLES IS DATA

3.2 OFFICIËLE STATISTIEKEN

3.2.1 Conceptenlijst

Om consistentie te waarborgen tussen statistieken en statistiekreeksen, hanteren we best een algemene conceptenlijst conform de SDMX-standaard. Deze conceptenlijst biedt een volledig overzicht van alle concepten die voorkomen in de door ons geproduceerde statistieken. Elk concept wordt voorzien van een korte identificatiecode, die kan worden gebruikt als snelle referentie in communicatie en als variabelenaam in datasets en datatabellen.

Grosso modo onderscheiden we twee soorten concepten:

- Algemene concepten: Dit zijn terugkerende concepten die vaak voorkomen in verschillende statistieken, zoals jaartal of geografisch gebied, maar ook zoals geslacht of NACE-code. De beschrijving van deze concepten is doorgaans beperkt.
- Het restconcept 'statistiekreeks': Dit concept omvat de rest van de operationele definities en dekt daarmee alle verdere details van een operationele definitie die niet onder de algemene concepten vallen.

Aan elk concept wordt bovendien een codelijst gekoppeld, eveneens conform de SDMX-standaard. Een codelijst bepaalt welke waarden een concept kan aannemen in een statistiekreeks. Voor het concept 'geslacht' bevat de codelijst bijvoorbeeld de waarden "man", "vrouw", en "andere", maar ook bijvoorbeeld de waarde "totaal", om totaalcijfers over beide geslachten aan te duiden.

Om consistentie te bewaren kunnen eveneens mastercodelijsten worden aangemaakt waaruit de codelijsten van gerelateerde concepten kunnen worden afgeleid. Zo kunnen we bijvoorbeeld de verschillende concepten "geslacht volgens rijksregister", "gerapporteerd geslacht" (in bevraging) en

“geaggregeerd geslacht” (voor geaggregeerde tabellen) definiëren waarvan de codelijsten sterk overlappen maar niet helemaal dezelfde zijn.

Voor het restconcept “statistiekreeks” kunnen de waarden zeer uitgebreid zijn. Bij statistiekreeksen die gebaseerd zijn op bevragingen omvat dit restconcept bijvoorbeeld in principe de volledige operationele beschrijving van de bevraging. Wanneer het ontwerp van de bevraging wordt aangepast, leidt dit tot een nieuwe statistiekreeks en een nieuwe waarde voor het restconcept. Net als bij de concepten voorzien we ook korte identificatiecodes voor elke waarde. Merk ook op dat de codelijst voor het concept “statistiekreeks” automatisch een overzicht biedt van alle statistiekreeksen die we produceren.

3.2.2 Hoe bepalen we concepten en dimensies?

De bepaling van statistiekreeksen op basis van concepten is geen exacte wetenschap, maar eerder een arbitrair proces waarin verschillende keuzes gemaakt moeten worden. Een eerste keuze is welke concepten je gebruikt als attributen om statistiekreeksen te onderscheiden en welke als dimensies binnen deze statistiekreeksen. In het ene uiterste gebruik je alle concepten als attributen waardoor een aparte reeks voor elke individuele statistiek ontstaat, zoals in tabel 6a. In het andere uiterste gebruik je alle concepten als dimensies waardoor één enkele reeks ontstaat met alle cijfers erin, zoals in tabel 6b. Beide benaderingen zijn uiteraard onpraktisch. We moeten op zoek naar een evenwicht dat zowel overzichtelijk als werkbaar is.

De tweede keuze is welke concepten binnen een statistiekreeks worden samengevoegd en welke worden uitgesplitst. Bijvoorbeeld: cijfers over mannen en vrouwen in gemeenten kun je uitsplitsen op basis van ‘parameter’ en ‘geslacht’, zoals in tabel 7b. Je kunt er echter ook voor kiezen om beide concepten te combineren in één dimensie, zoals in tabel 7a. Ook hier gaat het om een afweging tussen eenvoud en gebruiksgemak.

Een derde keuze is de definitie van de concepten zelf. Stel dat je cijfers publiceert over inwoners van Vlaamse gemeenten, uitgesplitst naar geslacht en leeftijd, zonder deze kenmerken te kruisen. Je kunt hiervoor aparte dimensies definiëren voor ‘geslacht’ en ‘leeftijd’, zoals in tabel 8a. Je kan deze dimensies echter ook herdefiniëren tot één dimensie ‘bevolkingsgroep’, die alle waarden van geslacht en leeftijd bevat, zoals in tabel 8b.

Deze voorbeelden tonen aan dat de bepaling van statistiekreeksen arbitrair is. Richtlijnen zijn daarom essentieel om consistentie en werkbaarheid te waarborgen.

Richtlijn 1. Cijfers die enkel verschillen in geografische indeling worden gecombineerd in één statistiekreeks waarbij deze geografische indeling als dimensie wordt opgevat. Bij voorkeur wordt voor zo’n geografische dimensie de NIS-code-indeling van Statbel gebruikt waarbij ook de versie

Tabel 6: Concepten kunnen worden gebruikt om aparte cijferreeksen te definiëren of als dimensie in één grote cijferreeks. Geen van beide situaties is echter werkbaar. We moeten een middenweg vinden.

(a) Elk cijfer in aparte statistiekreeks.

Aantal inwoners in Vlaanderen
6 821 770

Percentage Vlamingen met gevorderde digitale vaardigheden
26,1

Aantal zeugen in Vlaanderen
339 103

Ondergrens 95%-CI gemiddelde vertrouwen in provinciale overheid
2,34

...
...

(b) Één statistiekreeks voor alle cijfers.

Parameter	Cijfer
Aantal inwoners in Vlaanderen	6 821 770
Percentage Vlamingen met gevorderde digitale vaardigheden	26,1
Aantal zeugen in Vlaanderen	339 103
Ondergrens 95%-CI gemiddelde vertrouwen in provinciale overheid	2,34
...	...

Tabel 7: De indeling van dimensies is arbitrair. Kenmerken kunnen gecombineerd worden in één dimensie of als aparte dimensies worden opgenomen.

(a) Geslacht en parameter worden gecombineerd als één dimensie.

Bevolkingsaantal 2019		
Gemeente	Parameter	Observatie
A	totaal aantal	300
A	aantal vrouwen	180
A	percentage vrouwen	60
A	aantal mannen	120
A	percentage mannen	40
B	totaal aantal	500
B	aantal vrouwen	100
B	percentage vrouwen	20
B	aantal mannen	400
B	percentage mannen	80

(b) Geslacht en parameter zijn aparte dimensies.

Bevolkingsaantal 2019			
Gemeente	Geslacht	Parameter	Observatie
A	totaal	aantal	300
A	vrouwen	aantal	180
A	vrouwen	percentage	60
A	mannen	aantal	120
A	mannen	percentage	40
B	totaal	aantal	500
B	vrouwen	aantal	100
B	vrouwen	percentage	20
B	mannen	aantal	400
B	mannen	percentage	80

Tabel 8: Concepten in een statistiekreeks kunnen arbitrair worden geherdefinieerd.

(a) Aparte dimensies voor geslacht en leeftijd.

Bevolkingsaantal 2019			
Gemeente	Geslacht	Leeftijd	Observatie
A	mannen	totaal	90
A	vrouwen	totaal	130
A	totaal	18-30 jaar	60
A	totaal	31-65 jaar	120
A	totaal	66+ jaar	40
B	mannen	totaal	350
B	vrouwen	totaal	150
B	totaal	18-30 jaar	20
B	totaal	31-65 jaar	400
B	totaal	66+ jaar	80

(b) Geslacht en leeftijd worden gecombineerd in de dimensie 'bevolkingsgroep'.

Bevolkingsaantal 2019		
Gemeente	Bevolkingsgroep	Observatie
A	mannen	90
A	vrouwen	130
A	18-30 jarigen	60
A	31-65 jarigen	120
A	66+ jarigen	40
B	mannen	350
B	vrouwen	150
B	18-30 jarigen	20
B	31-65 jarigen	400
B	66+ jarigen	80

Tabel 9: Voor hiërarchische dimensies kunnen bijkomende dimensies worden toegevoegd, hoewel niet strikt noodzakelijk.

Bevolkingsaantal 2019			
Geografisch gebied	Gewest	Provincie	Aantal inwoners
Vlaams Gewest			6 589 069
Waals Gewest			3 633 795
Brussels Gewest			1 208 542
Prov. Antwerpen	Vlaams Gewest		1 857 986
Aartselaar	Vlaams Gewest	Prov. Antwerpen	14 293
Antwerpen	Vlaams Gewest	Prov. Antwerpen	525 935
Boechout	Vlaams Gewest	Prov. Antwerpen	13 266
Boom	Vlaams Gewest	Prov. Antwerpen	18 244
Borsbeek	Vlaams Gewest	Prov. Antwerpen	10 854
...

van de NIS-code-indeling vermeld wordt. Naast NIS-codes kunnen echter ook andere geografische gebieden, zoals postcodes, toeristische regio's of ziekenhuisnetwerken, worden opgenomen. Voor hiërarchische indelingen kunnen voor de volledigheid extra dimensies worden toegevoegd zoals in tabel 9.

Richtlijn 2. Cijfers die alleen verschillen in tijdsperiode worden samengevoegd in één statistiekreeks, met de tijdsperiode als dimensie.



De keuze om tijdsperiodes als dimensie te behandelen, betekent niet dat alle cijfers steeds in één dataset opgeslagen moeten worden. Zo is het op een file server handiger datasets per jaar te bewaren, terwijl deze datasets toch statistieken bevatten uit dezelfde statistiekreeks.

Richtlijn 3. Cijfers die op dezelfde manier worden berekend en uit dezelfde bron worden afgeleid, worden zoveel mogelijk gecombineerd in één reeks. Dit vereenvoudigt de verwerking en documentatie.

Richtlijn 4. Statistische parameters zoals frequenties en gemiddelden worden zo veel mogelijk opgesplitst in aparte reeksen. Dit vergemakkelijkt de beschrijving van de statistiekreeksen. Parameters die op een natuurlijke manier bij elkaar horen, vormen een uitzondering en worden wel zo veel mogelijk gecombineerd in dezelfde statistiekreeksen:

- Proporties/percentages berekend op frequenties uit één enkele statistiekreeks, worden ook aan deze statistiekreeks toegevoegd.
- Inferentiële statistieken zoals de onder- en bovengrens van betrouwbaarheidsintervallen, standaardfouten of p -waarden worden toegevoegd aan de statistiekreeksen met de bijhorende puntschattingen.



Percentages kunnen op verschillende manieren worden gedefinieerd. Dit moet altijd duidelijk worden omschreven in de operationele definitie en de parameterdimensie (bijv. het percentage mannen in gemeente X is niet hetzelfde als het percentage Vlaamse mannen die in gemeente X wonen).

Richtlijn 5. Binnen een statistiekreeks worden de verschillende dimensies zo goed mogelijk volledig met elkaar gekruist. Als dit niet mogelijk of wenselijk is, kunnen dimensies worden samengevoegd, bijvoorbeeld een dimensie “bevolkingsgroep” in plaats van aparte dimensies “geslacht”, “leeftijd” en “nationaliteit” als deze dimensies niet worden gekruist.

Richtlijn 6. Cijfers die gebaseerd zijn op meerdere basisreeksen afgeleid uit verschillende databronnen, zoals bevolkingsdichtheid (gebaseerd op inwonersaantallen en geografische oppervlakten), worden altijd in aparte reeksen ondergebracht. Dit bevordert overzichtelijkheid en documentatie.

Richtlijn 7. Alle concepten worden verzameld in een centrale conceptenlijst dat dient als datamodel. Dit datamodel wordt beheerd door een toegewezen team van collega's. Bij elk concept wordt ook een codelijst voorzien die alle mogelijke waarden op het concept bevat.

3.2.3 Versies van statistiekreeksen

Statistiekreeksen kunnen doorheen de tijd veranderen om verschillende redenen:

- De operationele definitie van een statistiekreeks wordt aangepast omdat dimensies veranderen, bijvoorbeeld de gemeentelideling verandert door gemeentefusies.

- De operationele definitie wordt aangepast in lijn met nieuwe gebruikersbehoeften.
- Er werd een fout ontdekt in de berekening van cijfers en deze fout wordt gecorrigeerd.

Omdat we verwacht worden elk gepubliceerd cijfer beschikbaar te houden, creëren we in elk van bovenstaande situaties een nieuwe versie van de statistiekreeks. Deze versie duiden we aan door een versienummer terwijl de cijfers en documentatie van de oude versie bewaard blijven. In het geval van een berekeningsfout voegen we een disclaimer toe aan de documentatie van de oude versie met uitleg over deze fout. Bij wijzigende definities, documenteren we eveneens waarom deze wijziging nodig was. Bij een nieuwe versie van een statistiekreeks publiceren we cijferreeksen ook retrospectief indien dat mogelijk is en gewenst wordt geacht.

4 ONTSLUITING STATISTIEKEN

We ontsluiten statistieken op twee manieren:

1. via volledige statistiekreeksen
2. via samenvattende cijferpagina's

Beide manieren worden hieronder besproken.

4.1 STATISTIEKREEKSEN

Een statistiekreeks is in essentie een dataset of datatabel die eenvoudig openbaar gepubliceerd kan worden. Het is mogelijk dat sommige cijfers in zo'n dataset een ontbrekende waarde hebben, bijvoorbeeld vanwege ontbrekende informatie of als gevolg van versluiering.

De structuur van een dataset met een statistiekreeks wordt bepaald volgens de SDMX-standaard. Deze standaard stelt dat een dataset met een statistiekreeks drie soorten informatie moet bevatten:

- Measures: Een kolom met de cijfers of statistieken. (In een toekomstige versie van SDMX kan een dataset meerdere kolommen met measures bevatten, waarbij het onderscheid tussen deze kolommen wordt vastgelegd als één van de dimensies).
- Dimensies: Kolommen met kenmerken van de cijfers die nodig zijn om elk cijfer uniek te identificeren.

- **Attributen:** Aanvullende informatie over de cijfers. Deze kan betrekking hebben op de gehele dataset, op een groep van cijfers of op één enkel cijfer.

De combinatie van alle dimensies en attributen bepaalt niet alleen de structuur van de data maar representeren samen ook altijd de volledige operationele definitie van de cijfers. Daarnaast heeft elke dataset een heldere en beknopte conceptuele definitie, oftewel een titel.

Het overzicht van de measures, dimensies en attributen wordt vastgelegd in de Data Structuur Definitie (DSD). De DSD bevat ook een verwijzing naar de codelijst van elke measure, dimensie en attribuut. Elke statistiekreeks heeft precies één DSD.

Bovendien moet elke statistiekreeks volledig vrij raadpleegbaar zijn. Ook oude versies van een statistiekreeks blijven beschikbaar. Deze datasets kunnen op verschillende manieren gepubliceerd worden, bijvoorbeeld:

- als een eenvoudige downloadbare dataset of een reeks van datasets (bijvoorbeeld in csv-bestandsformaat), of
- via een interactieve cijferapplicatie waarin gebruikers selecties kunnen maken op basis van dimensies en alleen relevante delen van een statistiekreeks kunnen visualiseren en downloaden.

4.2 CIJFERPAGINA'S

Cijferpagina's zijn webpagina's waarop specifieke statistieken in de schijnwerpers worden gezet. Het doel van een cijferpagina is niet om volledige statistiekreeksen te publiceren, maar om een toegankelijk en boeiend narratief rond enkele opvallende statistieken uit deze reeksen te vertellen. Voor de creatie van een cijferpagina worden er dus slechts enkele statistieken van een statistiekreeks geselecteerd om een helder en overzichtelijk verhaal te creëren dat aantrekkelijk is voor een breed publiek. Vanwege deze opzet kunnen niet alle beschikbare statistieken van een statistiekreeks op een cijferpagina worden besproken of gevisualiseerd. Dit zou de pagina immers onnodig complex en minder gebruiksvriendelijk maken.

Cijferpagina's bevatten vaak verschillende tabellen die een selectie van statistieken uit de statistiekreeksen presenteren. De structuur van deze tabellen is echter zelden uniform. Zo kan de eerste tabel bijvoorbeeld een tijdreeks met totaalcijfers tonen, waarbij 'tijdsperiode' de enige dimensie is. Een tweede tabel kan daarentegen cijfers weergeven voor slechts één enkel jaar, uitgesplitst naar verschillende bevolkingsgroepen. In dat geval is er gefilterd op één tijdsperiode en vormt 'bevolkingsgroep' de dimensie. Deze variatie in tabelopmaak maakt duidelijk dat cijferpagina's niet bedoeld zijn om volledige statistiekreeksen te ontsluiten.

De inhoud van een cijferpagina wordt samengesteld door een team van collega's verantwoordelijk voor de ontsluiting van onze statistieken op een toegankelijke manier. Dit proces staat los van de productie van de statistiekreeksen zelf, wat ervoor zorgt dat de focus blijft liggen op de creatie van begrijpelijke en informatieve verhalen voor een breed publiek.

Ideeën voor verdere ontwikkeling van deze handleiding

- Handleiding over hoe je een duidelijke operationele definitie schrijft, opgedeeld in de verschillende concepten?
- Diepgaandere uitwerking van definitie gebruikersbehoeften?
- Handleiding SDMX?
- Handleiding SIMS/GSBPM/...?
- Handleiding over hoe kernstatistieken worden aangeleverd aan de VSA (als een gefilterde statistiekreeks).
- ...?