STATISTIEK
VLAANDEREN

FOLDERSTRUCTUUR VSA

10.06.2025

Vlaanderen
is sterk in cijfers

Deze nota vertrekt vanuit de nota van Noëmi:
VSA_Fileserver_Structuur_Werkdocument.docx.

Nog openstaande vragen:

- *Waar worden dataprocessen R&D-gewijs ontwikkeld? -> In de themafolders?*
- *Waar gebeuren eenmalige analyses, zoals de aanmaak van rapporten? -> Nieuw thema aanmaken voor rapporten, analoog aan TTP?*
- *Wat met safe-roomactiviteiten? We kunnen externe gebruikers niet zomaar toegang geven tot alle data die onder collega's wel gedeeld wordt. -> Nieuw thema, analoog aan TTP?*
- *Waar plaatsen we analyses van helpdeskvragen? -> Nieuw thema, analoog aan TTP?*

We gebruiken de fileserver om bestanden centraal op te slaan en eenvoudig te delen met collega's. De aard van deze bestanden is erg uiteenlopend. In de eerste plaats bewaren we er datasets. Deze variëren sterk in formaat, structuur en beveiligingsvereisten. Sommige bestanden zijn bronbestanden die rechtstreeks afkomstig zijn van dataleveranciers; andere datasets worden door onszelf aangemaakt of bewerkt.

Daarnaast gebruiken we de fileserver om dataverwerkingsscripts op te slaan. Deze scripts sturen verschillende programma's aan, zoals R, Python, Pentaho, SPSS, Markdown of LaTeX.

Tot slot kan de fileserver ook documentatie bevatten. Het gaat dan vooral om documentatie die automatisch wordt gegenereerd via scripts, zoals de responsanalyses van de SV-bevragingen. Voor andere soorten documentatie is het belangrijk af te wegen of deze niet beter thuishoren op SharePoint.

***Discussiepunt**: Het lijkt logischer om interne procesdocumentatie op de fileserver te bewaren in plaats van op SharePoint, zodat deze dicht bij de bijhorende scripts staat. Akkoord?*

***Discussiepunt**: Data kan ook in een database worden opgeslagen. We zullen duidelijke afspraken moeten maken over welke data waar thuishoort. Bronbestanden die fysiek worden aangeleverd, moeten sowieso op de fileserver staan. Bewerkte datasets kunnen in principe ook in een database worden geplaatst. Wat in de praktijk het meest efficiënt*

is, is momenteel nog niet helemaal duidelijk. Databases zijn vaak efficiënter voor dataverwerking, onder andere omdat data niet volledig in het geheugen hoeven te worden geladen. Anderzijds zijn de bewerkingsmogelijkheden in een database soms beperkt, en kunnen parquet-bestanden op de fileserver tegenwoordig ook al zeer efficiënt worden opgeslagen en verwerkt.

Deze nota vat samen hoe we de fileserver organiseren en licht de gemaakte keuzes toe. Op die manier willen we meer overzicht creëren over alle bestanden en datasets die we op de fileserver bewaren.

1 ALGEMENE AFSPRAKEN

Voor we in de opdeling van de fileserver duiken, maken we eerst een algemene afspraak over hoe we onze folders namen geven.

- Vermijd spaties. Spaties geven vaak problemen in scripts of command-line tools. Gebruik liever `-` of `_`. Bijvoorbeeld: `project_2024/` of `data-cleaned/`.
- Begin met een datum voor chronologische sortering, gebruik daarbij het ISO-formaat `YYYYMMDD`. Bijvoorbeeld: `20250424_meeting-notes/`.
- Nummer versies of iteraties expliciet. Bijvoorbeeld: `v1/`, `v2/`, `final_v3/`, `01_raw/`, `02_cleaned/`.
- Wees beschrijvend maar beknopt. Vermijd vage namen zoals `data1/`. Bijvoorbeeld: `raw_data/`, `thesis_figures/`.
- Gebruik een consistente structuur. Kies één stijl en hou je eraan: `kebab-case` (`my-folder`), `snake_case` (`my_folder`), of `camelCase` (`myFolder`).
  Hier moeten we één standaard kiezen voor VSA. Welke nemen we? In R is `snake_case` het meest gebruikelijk, maar we hoeven dit niet noodzakelijk over te nemen voor onze folderstructuur.
- Vermijd speciale tekens en accenten zoals: `&`, `@`, `!`, `é`, `ç`, enzovoort. Deze kunnen problemen veroorzaken bij synchronisatie of in andere systemen.
- Begin niet zomaar met cijfers, tenzij dit dient voor de ordening van mappen. Bijvoorbeeld: `01_code/`, `02_data/`.
- Gebruik bij voorkeur Engelse termen in plaats van Nederlandse. Dit verkleint het risico op taalproblemen, bijvoorbeeld bij communicatie met Eurostat of samenwerking met niet-Nederlandstalige consultants.
  Iedereen akkoord?

2 OPDELING VAN DE FILE SERVER

Om onze fileserver overzichtelijk te organiseren, structureren we de bestanden in mappen volgens een duidelijke hiërarchie. Hieronder lichten we de opbouw toe.

2.1 LAAG 1: INHOUDELIJK THEMA

Binnen de VSA werken we rond verschillende inhoudelijke thema's, zoals *bevolking*, *bevragingen*, *ondernemingen* of *toerisme*. Elk thema wordt doorgaans opgevolgd door een stabiele groep collega's. Het is dan ook logisch om de bestanden op de fileserver te organiseren volgens deze thematische indeling. Dit vormt het eerste mappenniveau.

De belangrijkste vraag is hoe we de thema's precies afbakenen. Zo'n afbakening is vaak arbitrair, maar essentieel om efficiënt samen te werken, werk te verdelen en databeveiliging te waarborgen. We hanteren hiervoor de volgende principes:

- Elk thema wordt beheerd door een stabiele, duidelijk omschreven groep collega's: de *themabeheerders*. Dit vereenvoudigt en verduidelijkt het rechtenbeheer.
    - Binnen elk thema dragen alle themabeheerders samen de verantwoordelijkheid voor de dataprocessen. Zij hebben kennis van elkaars werk, volgen dit op en evalueren het. Daarom mag een thema niet te smal (één persoon) of te breed (te veel personen) zijn gedefinieerd.
    - Collega's kunnen aan meerdere thema's meewerken, maar we beperken dit zoveel mogelijk en streven naar een logische indeling.
    - Elk thema wordt opgevolgd door minstens twee inhoudelijke themabeheerders, om wederzijdse kwaliteitscontrole te garanderen en risico's bij uitval te beperken.
- Thema's volgen zoveel mogelijk de logica van het dataverwerkingsproces: van brondatabestand tot afgewerkt product gebeurt de verwerking idealiter binnen één themagroep.
    - Onder afgewerkte producten verstaan we niet alleen Vlaamse Openbare Statistieken, maar ook GSM-indicatoren, PUF's, SUF's, SV-rapporten, enzovoort.
    - Dit sluit niet uit dat gegevens uit één thema ook in andere thema's worden gebruikt. Bijvoorbeeld: bevolkingsdata behoren tot het thema *demografie*, waarbij alleen de themabeheerders toegang hebben tot de atomaire data. De geaggregeerde cijfers worden vervolgens beschikbaar gesteld voor het hele VSA-team, zodat ze

ook in andere thema's kunnen worden gebruikt — bijvoorbeeld als noemers voor GSM-indicatoren of voor de berekening van gewichten in de SV-bevraging.

- Thema's kunnen zowel recurrente dataleveringen (jaarlijks, maandelijks, ...) als eenmalige ad-hocleveringen (zoals TTP-activiteiten) bevatten.
- De thematische indeling moet altijd voldoen aan de beveiligingsafspraken met dataleveranciers. Collega's die volgens deze afspraken geen toegang mogen hebben tot bepaalde data, kunnen ook geen themabeheerder zijn voor dat thema. Bijvoorbeeld: als met Statbel is overeengekomen dat enkel Karolien toegang heeft tot toeristische overnachtingsdata, wordt dit een apart thema — ook al werken andere collega's aan gerelateerde onderwerpen binnen een ander thema *toerisme*. (We moeten dan natuurlijk ook evalueren of het verstandig is om deze verantwoordelijkheid bij slechts één persoon te leggen, zie het eerste punt.)

- Thema's kunnen worden gegroepeerd in bredere *themagroepen*.
  - Elk TTP-project vormt een apart thema met eigen data en een beperkte groep collega's die toegang hebben. Al deze projecten worden ondergebracht in de folder 'TTP'.
  - Andere thema's kunnen worden samengevoegd tot een themagroep wanneer de inhoudelijke samenhang groot is. Zo kunnen de VOS-thema's *Buitenlandse handel*, *Economie* en *Financiële omgevingsfactoren* worden samengebracht onder de themagroep *Economie*.
- Naast de inhoudelijke thema's zijn er nog twee folders die op dit niveau thuishoren. Ze volgen dezelfde logica, maar zijn strikt genomen geen thema's of themagroepen:
  - **R-package**: Deze map bevat een R-pakket met functies die het programmeerwerk binnen andere thema's vereenvoudigen. De ontwikkeling gebeurt door een kleine groep collega's met doorgedreven R-kennis.

    Vraag: *Moet dit niet een algemene folder zijn voor allerlei hulpprogramma's, waarbij het R-pakket slechts één subfolder is?*
  - **Datamodel**: Deze map bevat het algemene datamodel dat door alle thema's heen kan worden gebruikt. Ook dit model wordt beheerd door een specifiek aangeduid team van collega's.

Hieronder volgt een overzicht van alle thema's en themagroepen. Bij elk thema wordt opgelijst welke collega's verantwoordelijk zijn, welke data met beperkte

toegang worden binnengehaald, welke open data worden verwerkt, en welke producten worden opgeleverd.

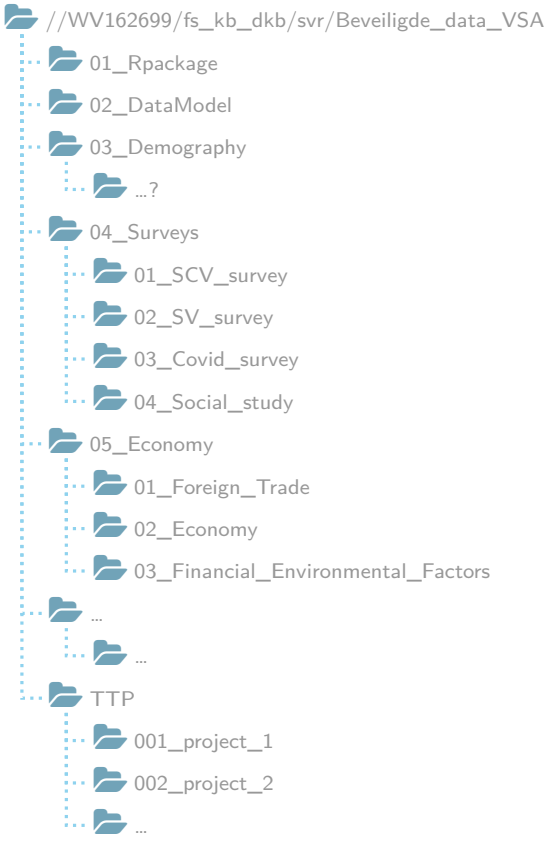
Discussiepunt: *Welke thema's bakenen we concreet af? Wie doet een voorstel?*

- → *Binnen het DWH-project werden zes thema's/themagroepen onderscheiden: Bevolking, Samenleving, Economie, Ruimte, Overheid en SV-survey. Ik heb echter nog geen motivatie voor deze indeling gezien en vraag me af of ze wel het meest efficiënt is. De thema's zijn zeer breed en de teams vaak erg heterogeen samengesteld. In het bestand Medewerkers_Activiteiten_Data.xlsx staat een verouderd overzicht van de collega's, thema's en beveiligde data.*
- → *Een andere mogelijke indeling is die van de VOS'en in thema's en subthema's (zie VSP-lijst_finaal_bestaande_VOS_aanpassingen.xlsx), al is deze waarschijnlijk te fijnmazig.*
- → *Een thema zoals demografie wordt momenteel door een kleine, goed afgebakende ploeg beheerd. Binnen dit thema worden echter veel verschillende datasets verwerkt, zoals stand en loop van de bevolking, censusdata, enzovoort. We kunnen deze processen ook als aparte thema's beschouwen om het overzicht te verbeteren, ook al zijn de verantwoordelijken dezelfde.*
- → *Heel wat output, zoals de GSM-indicatoren, is momenteel nog niet duidelijk gekoppeld aan een thema (of ik heb daarover nog geen informatie gevonden). Klopt dat?*

Opdracht: *Ik heb hieronder alvast een voorbeeld uitgewerkt voor het thema demografie (als we beslissen om dit als één thema te blijven zien). De lijst moet worden aangevuld voor alle andere thema's. Jo en Lisa nemen dit op.*

Thema	Beheerders	Data-Inname Beperkte Toegang	Data-Inname Open Toegang	Producten
Demography	Ingrid Schockaert Jan Pickery Lisa Van Landschoot	Demobel (Statbel); Oekraïense vluchtelingenbevolking (Statbel); Aanwezigen register tijdelijke bescherming (DVZ); Census (Statbel); Toerisme (Statbel)	geen?	VOS Bevolking (101, 102, 103, 104, 105); VOS Migratie (106, 107, 108); VOS Geboorte en sterfte (109, 110, 331, 332); VOS Relaties en huishoudens (111, 112, 113, 114); VOS Vooruitzichten (115, 554); GSM-indicatoren DE_00, DE_01, DE_02, DE_03, DE_04, DE_05, DE_06, DE_07, DE_08, DE_09, DE_10, DE_12, DE_15, DE_16, DE_17, DE_18, DE_19, DE_20, DE_21, DE_22, IN_02, IN_03, DE_25, DE_26, DE_27, IN_04, IN_05, DE_31, DE_32, DE_33, DE_34, DE_35, DE_36, IN_06, IN_07, IN_08, DE_40, IN_09, DE_42, IN_10, IN_11, IN_12, IN_13, IN_14, IN_15, DE_64, DE_66, DE_69, DE_74, DE_75, DE_76, DE_77, DE_78, DE_79, IN_30, IN_31, IN_32; aantal inwoners voor intern gebruik binnen VSA volgens geslacht, leeftijd,...; ...?
...	...	...	...	...
TTP/Project001	data steward data engineers projectmedewerkers	elke binnenkomende dataset	per definitie geen	gevraagd product van TTP-project
TTP/Project...	data steward data engineers projectmedewerkers	elke binnenkomende dataset	per definitie geen	gevraagd product van TTP-project

De eerste laag van de folderstructuur zal er dus ongeveer als volgt uitzien:



Discussiepunt: *Momenteel worden alle kleinere geaggregeerde datasets samen verzameld in de folder 0101_Geaggregeerde_data_Collect. Deze datasets hoeven niet verder beveiligd te worden en zijn voor alle collega's toegankelijk. Beveiligde datasets worden echter nog steeds in aparte folders bewaard in 4_Beveiligde_data met beperkte toegang voor sommige collega's. We werken dus met twee systemen. Houden we dit zo? Dit lijkt niet echt efficiënt. Voer voor discussie? Op basis van mijn voorstel voor de volgende mappenlaag, gaan we steeds gemakkelijk een overzicht kunnen krijgen van alle brondatabestanden.*

2.2 LAAG 2: FUNCTIE EN BEVEILIGING

Naast de inhoudelijke thema's delen we bestanden ook op volgens de functie van het bestand en het beveiligingsniveau. Wat betreft de functie van bestanden maken we ten eerste het volgend onderscheid:

- datasets (zowel inkomende als verwerkte),
- scripts om datasets te verzamelen en te verwerken, en
- documentatie over de scripts en de data.

(Merk op dat scripts ook kleine databestanden kunnen bevatten die op natuurlijke wijze bij een script horen, bijvoorbeeld een mapping om één variabele om te zetten naar een andere. Dit is geen "inhoudelijke" data.)

Wat betreft beveiliging maken we ten tweede een onderscheid tussen verschillende beveiligingsniveaus:

- Bestanden met beperkte toegang. Dit zijn bestanden die in principe niet aangepast mogen worden. Hierbij gaat het voornamelijk over data die we binnenhalen. Zo'n data worden opgeslagen en nadien nooit meer aangepast.
- Bestanden met geprivilegieerde toegang. Dit zijn bestanden die enkel toegankelijk zijn voor themabeheerders zoals bijvoorbeeld de gekuiste versies van vertrouwelijke brondata.
- Interne bestanden: Deze bestanden zijn raadpleegbaar door iedereen binnen de VSA, maar zijn niet bedoeld om breder publiek te verspreiden. Deze bestanden kunnen worden gebruikt als bron in andere dataprocessen. Bijvoorbeeld, een dataset met het aantal inwoners in elke gemeente wordt als intern bestand beschikbaar gesteld door de demografen zodat andere collega's deze kunnen gebruiken als noemers bij de berekening van GSM-indicatoren.
- Publieke bestanden: Dit zijn bestanden die in principe vrij beschikbaar mogen zijn voor het brede publiek. Hier vallen bijvoorbeeld alle datasets onder die worden gebruikt voor de cijferapplicaties en cijferpagina's. Net zoals bij interne bestanden gelden hier lees- en schrijfrechten voor de themabeheerders, en leesrechten voor alle andere VSA-medewerkers.

De opdeling in beveiligingsniveaus geldt enkel voor datasets. We volgen het algemene principe dat scripts en documentatie van onze dataprocessen steeds toegankelijk zijn voor alle collega's. Dit verhoogt transparantie binnen het team en bevordert ook dat we van elkaar leren.

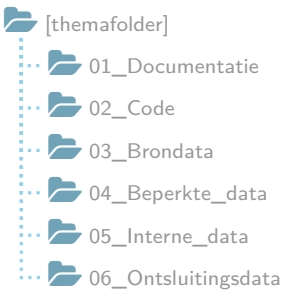
Omdat documentatie en scripts steeds interne bestanden zijn en enkel datasets verder worden opgedeeld volgens beveiliging, is het overzichtelijker om beide opdelingen samen te nemen in één folderlaag. We gebruiken binnen deze laag de volgende opdeling:

- Folder '`01_documentation`' wordt gebruikt om alle documentatie over een dataproces te bewaren. Deze documentatie wordt geschreven en beheerd door de themabeheerders. Binnen deze folder gelden daarom enkel lees- en schrijfrechten voor de data-engineers en de themabeheerders. Alle andere VSA-collega's hebben enkel leesrechten.
- Folder '`02_code`' bevat alle scripts om data te verwerken. Net zoals de documentatie wordt dit beheerd door de themabeheerders. Binnen deze folder gelden daarom enkel lees- en schrijfrechten voor de data-engineers en de themabeheerders. Alle andere VSA-collega's hebben enkel leesrechten.

  We gebruiken een aparte folder zodat we dit kunnen gebruiken voor git.

- Folder '03_source_data' bevat datasets zoals ze worden aangeleverd door dataleveranciers. Deze datasets worden enkel bewaard in deze folder maar nooit aangepast. Het is de taak van de themabeheerders om deze brondata op te kuisen en af te toetsen aan het datamodel. Nadien bewaren zij de gekuiste data in één van de volgende drie folders, afhankelijk van het beveiligingsniveau en de functie van de gekuiste data. In deze folder gelden enkel schrijfrechten voor de data-engineers die als enigen de bestanden kunnen opslaan. Er gelden leesrechten voor de themabeheerders zodat ze de data kunnen aanroepen voor verdere verwerking. Geen enkele andere collega heeft toegangsrechten tot deze folder.
- Folder '04_privileged_data' bevat verwerkte data die niet toegankelijk mogen zijn voor andere collega's. Het gaat hierbij om data waarin persoonlijke informatie zit of gegevens die nog versluierd moeten worden, bijvoorbeeld de gekuiste en gevalideerde versies van persoonsdata. Op deze bestanden gelden dan ook enkel lees- en schrijfrechten voor de themabeheerders en de data-engineers. Geen enkele andere collega heeft toegangsrechten tot deze folder.
- Folder '05_internal_data' bevat data die wel toegankelijk zijn voor collega's maar niet de bedoeling hebben verder te worden verspreid onder het brede publiek. Onder deze folder kunnen bestanden vallen met onversluierde data waarvoor we wel de toestemming krijgen deze binnen de VSA te verspreiden. Scientific use files (SUF's) vallen onder deze categorie. Deze bestanden worden beheerd door de themabeheerders die er lees- en schrijfrechten op hebben. Andere collega's hebben leesrechten zodat ze de bestanden kunnen aanroepen, maar ze hebben geen schrijfrechten.
- Folder '06_public_data' bevat bestanden die publiek verspreid mogen worden. Het gaat hierbij bijvoorbeeld om bestanden voor de cijferapplicaties en cijferpagina's of public use files (PUF's). Versluierde gevoelige data kunnen hier ook hun onderdak vinden. Deze bestanden worden eveneens beheerd door de themabeheerders die er lees- en schrijfrechten op hebben. Andere collega's hebben enkel leesrechten zodat ze de bestanden kunnen aanroepen, maar ze hebben geen schrijfrechten.

Kortom, onder elke themafolder kunnen enkel de volgende folders worden gebruikt:

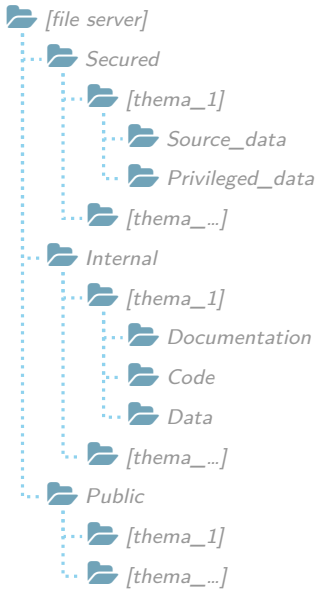


Een overzicht van de toegangsrechten staat beschreven in de volgende tabel:

	Schrijfrechten	Leesrechten	Geen toegang
01_Documentatie	Themabeheerders Data-engineers	Alle collega's	—
02_Code	Themabeheerders Data-engineers	Alle collega's	—
03_Brondata	Data-engineers	Themabeheerders Data-engineers	Alle andere collega's
04_Beperkte_data	Themabeheerders Data-engineers	Themabeheerders Data-engineers	Alle andere collega's
05_Interne_data	Themabeheerders Data-engineers	Alle collega's	—
06_Ontsluitingsdata	Themabeheerders Data-engineers	Alle collega's	—

Vraag aan Lieven en Georneth: *Is deze organisatie haalbaar voor jullie op het vlak van rechtenbeheer? Zijn dit niet te veel mappen waarvoor verschillende rechtenregimes moeten worden ingesteld? Kan rechtenbeheer geautomatiseerd worden op basis van een overzichtsbestand?*

Opmerking/vraag: *Met dit voorstel voor folderstructuur kunnen we gemakkelijk migreren naar een structuur met meerdere servers. Stel dat we in de toekomst een aparte server opzetten met de publieke data, hoeven we enkel over alle folders* `public_data` *te itereren en de inhoud te migreren.*
Een alternatief is dat we zo'n serverstructuur al op het eerste niveau implementeren van de folderindeling implementeren, maar dat leidt tot een complexere structuur.
Bijvoorbeeld:



Probleem is dat we op dit moment nog geen beslissing hebben over welke servers we in de toekomst apart gaan installeren voor data met verschillende beveiligingsniveaus.

2.3 LAAG 3: PROCESSPECIFIEK

Vrij in te richten door themabeheerders, kleine processen hebben minder structuur nodig dan grote processen.

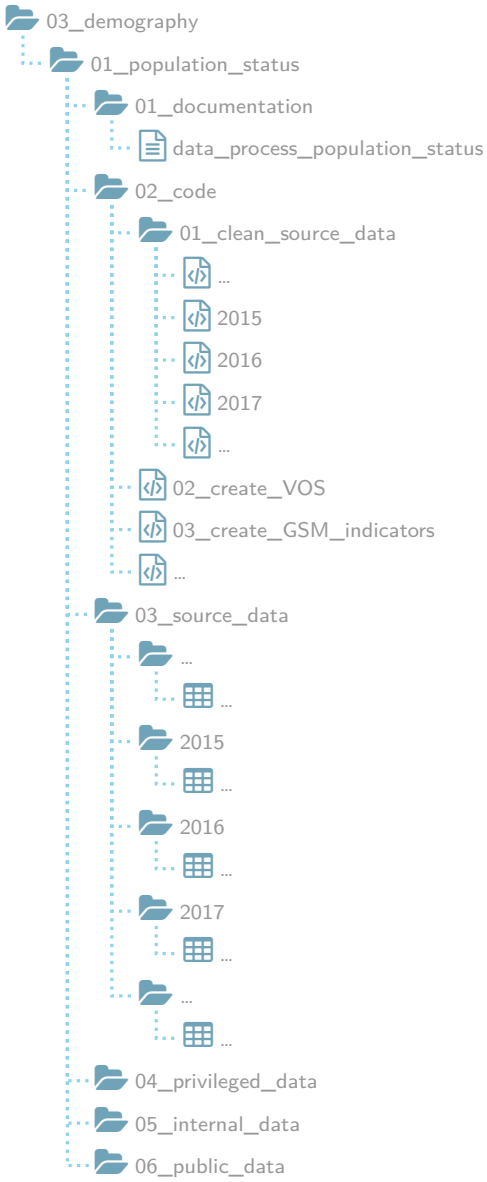
GSBPM is leidraad, maar moet niet strikt worden opgevolgd (zoals ook staat beschreven in documentatie van GSBPM zelf). Daarvoor zijn processen te divers. Voor sommige processen is GSBPM overkill, voor anderen is de volgorde van stappen niet meest efficiënt.

Good practice om folders te ordenen volgens verwerkingsproces, voorbeeld toevoegen

Hier algemene afspraken toevoegen ipv aan het begin van het document?

Richtlijnen toevoegen voor opslag brondata, eerst volgens referentieperiode, dan volgens versie.

Een voorbeeld voor demografie — stand van de bevolking —.


03_demography
01_population_status
01_documentation
data_process_population_status
02_code
01_clean_source_data
...
2015
2016
2017
...
02_create_VOS
03_create_GSM_indicators
...
03_source_data
...
...
2015
...
2016
...
2017
...
...
...
04_privileged_data
05_internal_data
06_public_data