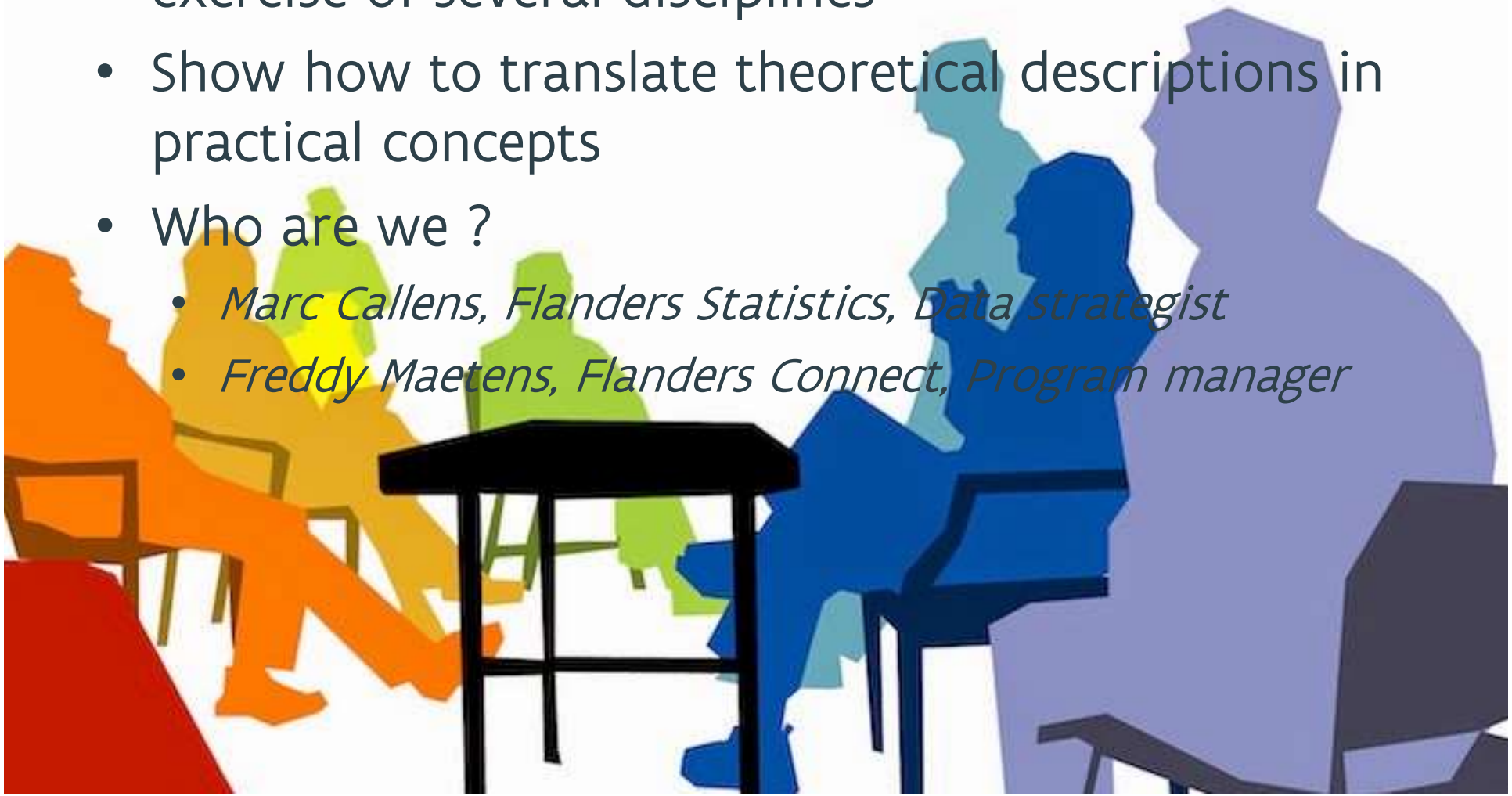
STATISTIEK
VLAANDEREN

Defining
a practical approach to realize a Statistical Datawarehouse platform Using GAMSO GSBPM GSIM

Vlaamse overheid

Objectives of this presentation

- Demonstrate that creating a S-DWH is a combined exercise of several disciplines
- Show how to translate theoretical descriptions in practical concepts
- Who are we ?
  - *Marc Callens, Flanders Statistics, Data strategist*
  - *Freddy Maetens, Flanders Connect, Program manager*



Short introduction to Statistics Flanders

STATISTIEK
VLAANDEREN

Vlaamse overheid

STATISTIEK
VLAANDEREN

Vlaamse overheid

Context

IIS

Interfederal Statistical Institute

Federal

STATBEL

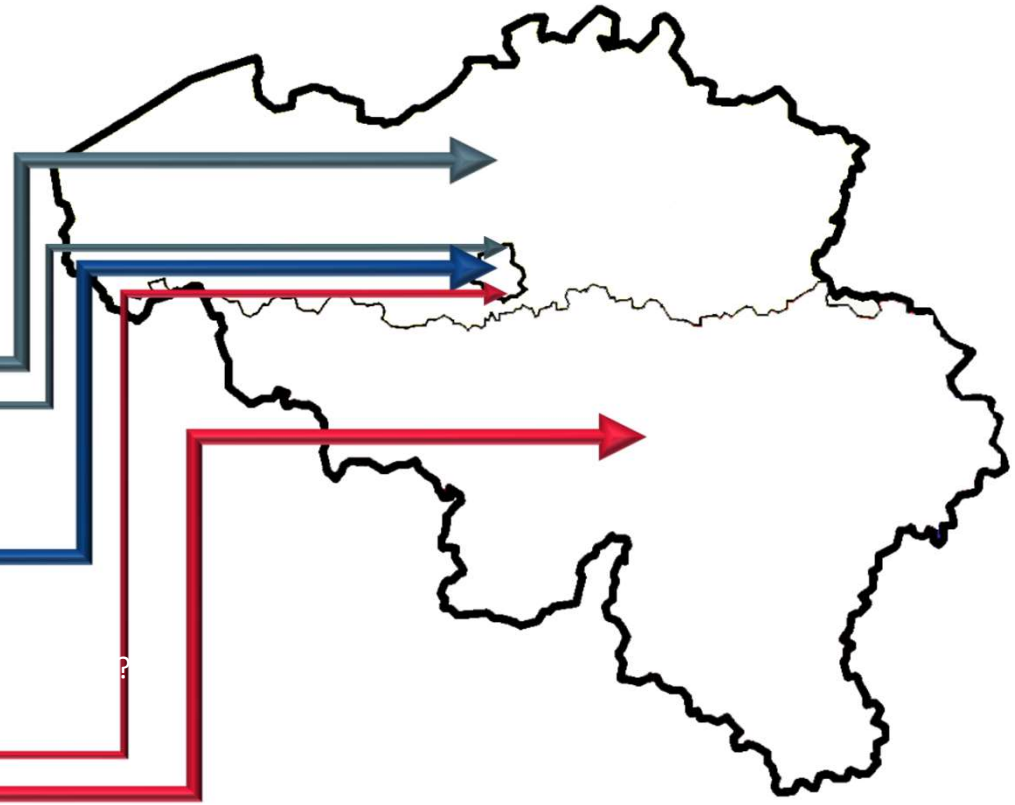
Regions / Communities

STATISTIEK VLAANDEREN

BISA Brussels Instituut voor Statistiek en Analyse

Wallonie

Iweps



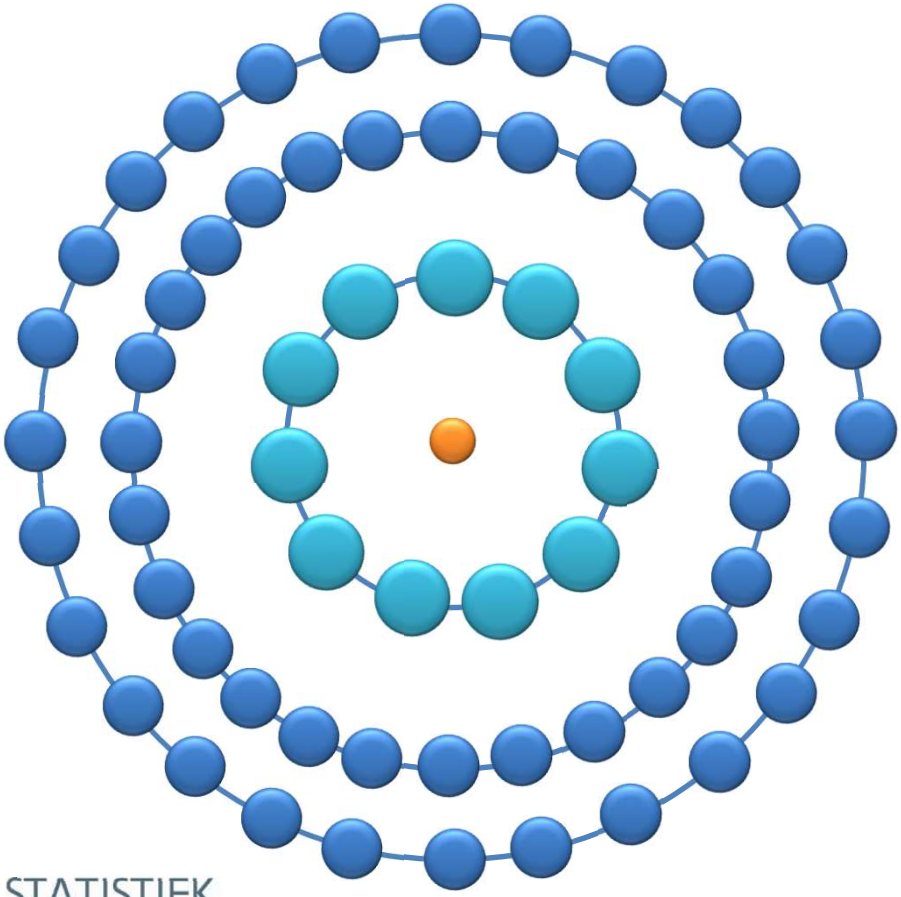
STATISTIEK VLAANDEREN

Simplified representation of the Belgian statistical landscape.

Vlaamse overheid

Coordination of a decentralized network

Network Statistics Flanders



FSA

Policy areas

Entities

Decentralized production between:

- 11 policy areas
- 55 entities
- VGC: Flemish Community Commission
- Municipalities & provinces

Sources:

- Regional, local and federal sources and data owners
- Also sources from outside the government

STATISTIEK VLAANDEREN

Vlaamse overheid

Flemish statistical authority

- 3 Objectives:
  - Coordination in collaboration with
    - Council for Flemish Public Statistics (RVOS)
    - Operational forum for Flemish Public Statistics (CVOS)
  - Production:
    - Official Flemish statistics (VOS),
    - Other output
      - in depth analysis,
      - experimental statistics
      - methodology
  - Data & Methods:
    - Mainly supporting the other 2 objectives

Introduction to Statistical Data Warehouse

Statistical Datawarehouse

Introduction to the S-DWH Project - Objectives

- Modernisation of the **statistical methodology** by introducing new methods, techniques and tools
- **ICT modernization** of the statistical production process; collect, store, analyse and disseminate statistical data
- **Increase efficiency** by
  - Linking of data
  - Re-utilization of data
- Create a platform to organize data exchanges with the network and with federal statistical institute

S-DWH Organisational constraints - challenges

- As-Is situation
  - Due to historical reasons know how and approach is very dispersed
- Limited resources
- Infrastructural landscape at Flemish Government
  - IT infrastructure and access management is managed centrally => Has to be brought in line with statistical legislation and approach

Starting point: COE Roadmap to build S-DWH

Start Project

- Business Requirements
- Establish Project Target
- Define Project Strategy
- Cost-benefit Analysis

Approved Business Case

- Building Blocks
- Business Architecture
- Information Architecture
- Metadata Model
- Data Linking
- Data Cleaning
- Estimation

Approved Design

- Technology Architecture
- Workflow System
- Metadata System
- Test

Working S-DWH

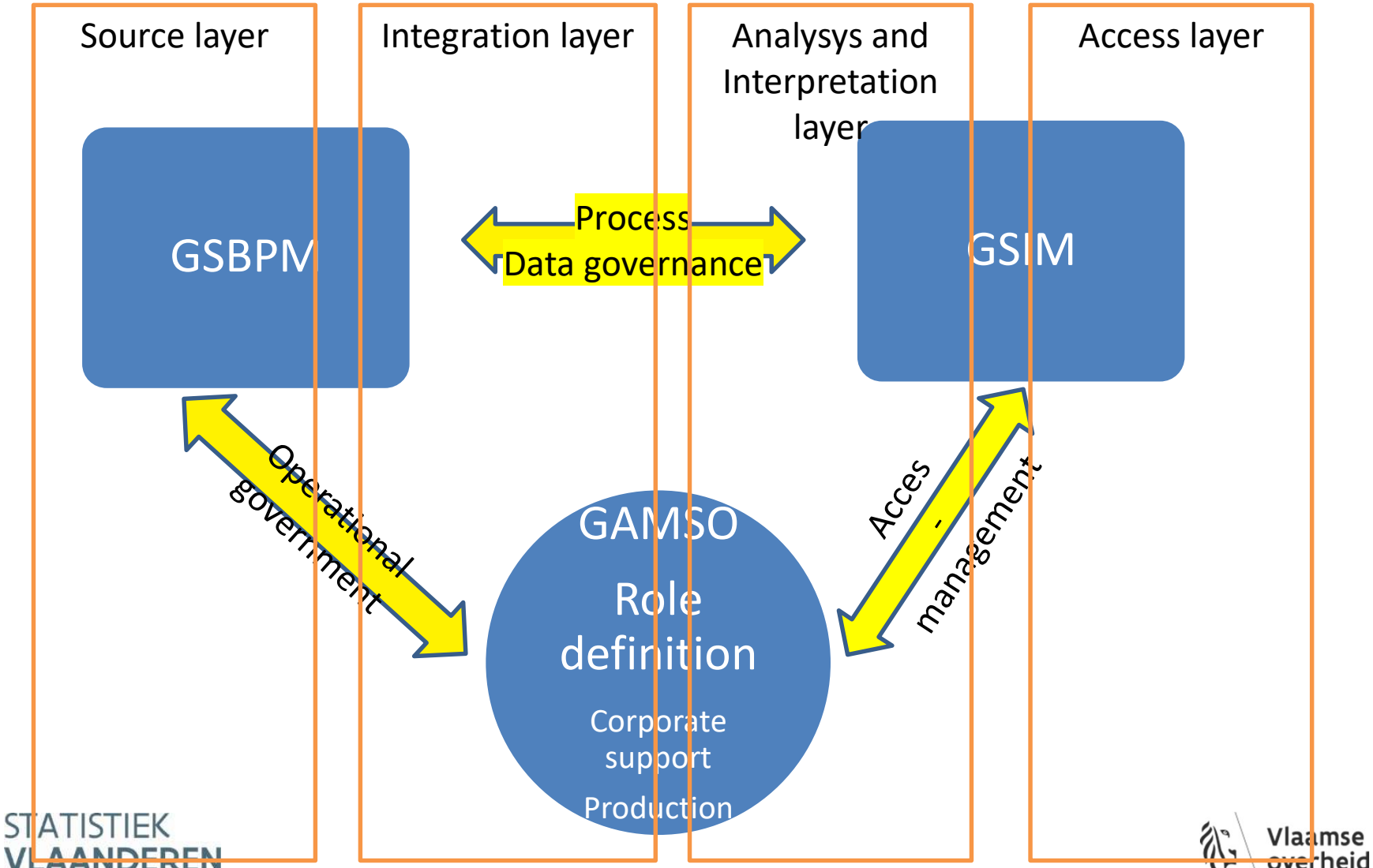
- Metadata Governance
- Confidentiality
- Analysts
- Revisions
- Training

Operational S-DWH

Theoretical phase

- How will we link the different methodologies to each other?

Theoretical phase : link the methodologies



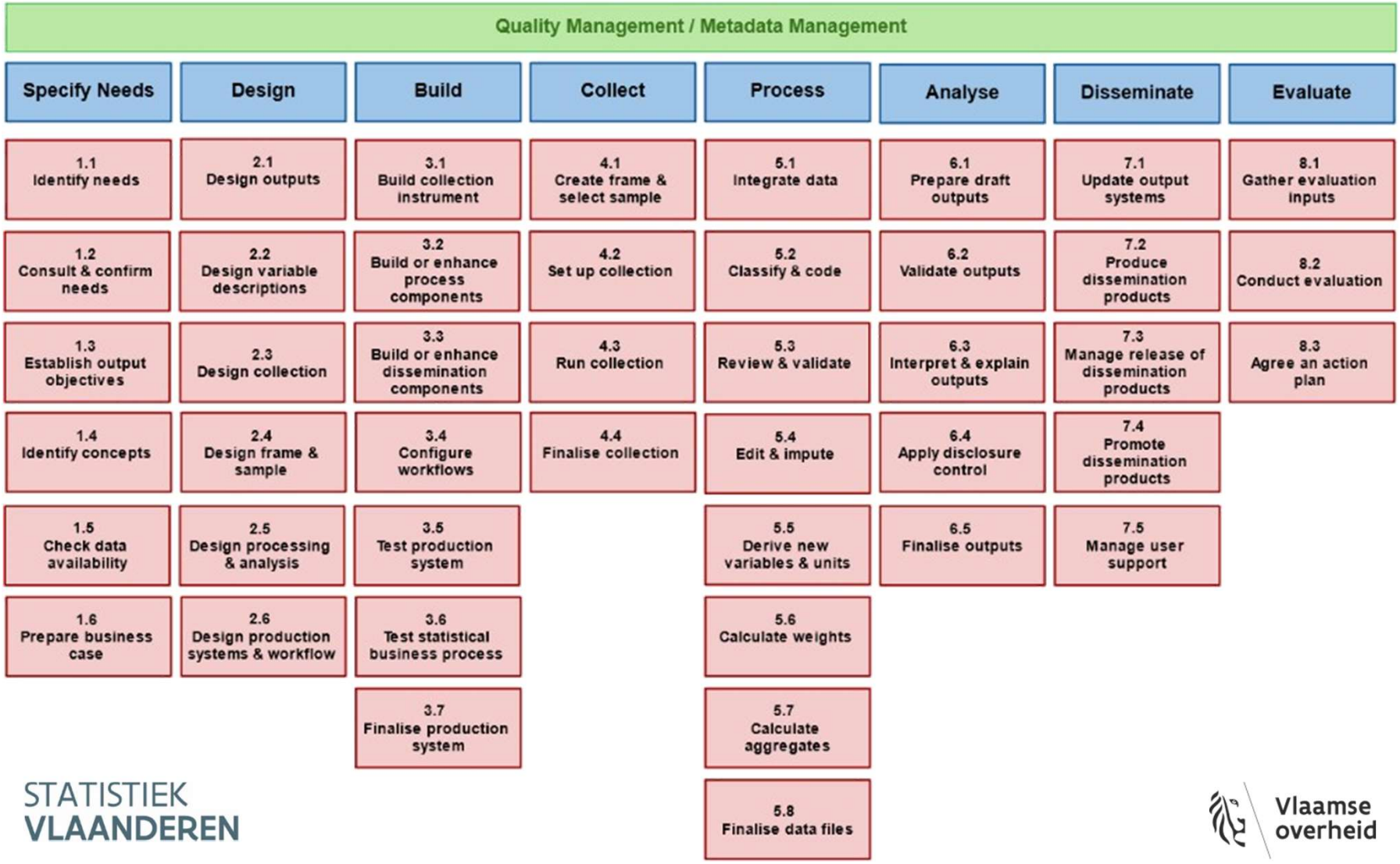
STATISTIEK VLAANDEREN

Vlaamse overheid

GAMSO provides a framework

- Production roles
  - Linked to GSBPM phases 4 to 7
- Corporate support
  - Linked to all GSBPM phases : methodology, broader support
- Strategy & leadership
  - Linked to a 5year strategy plan which gave us long term objectives
- Capability management
  - Linked to different work groups defining and evaluation competences, training and education

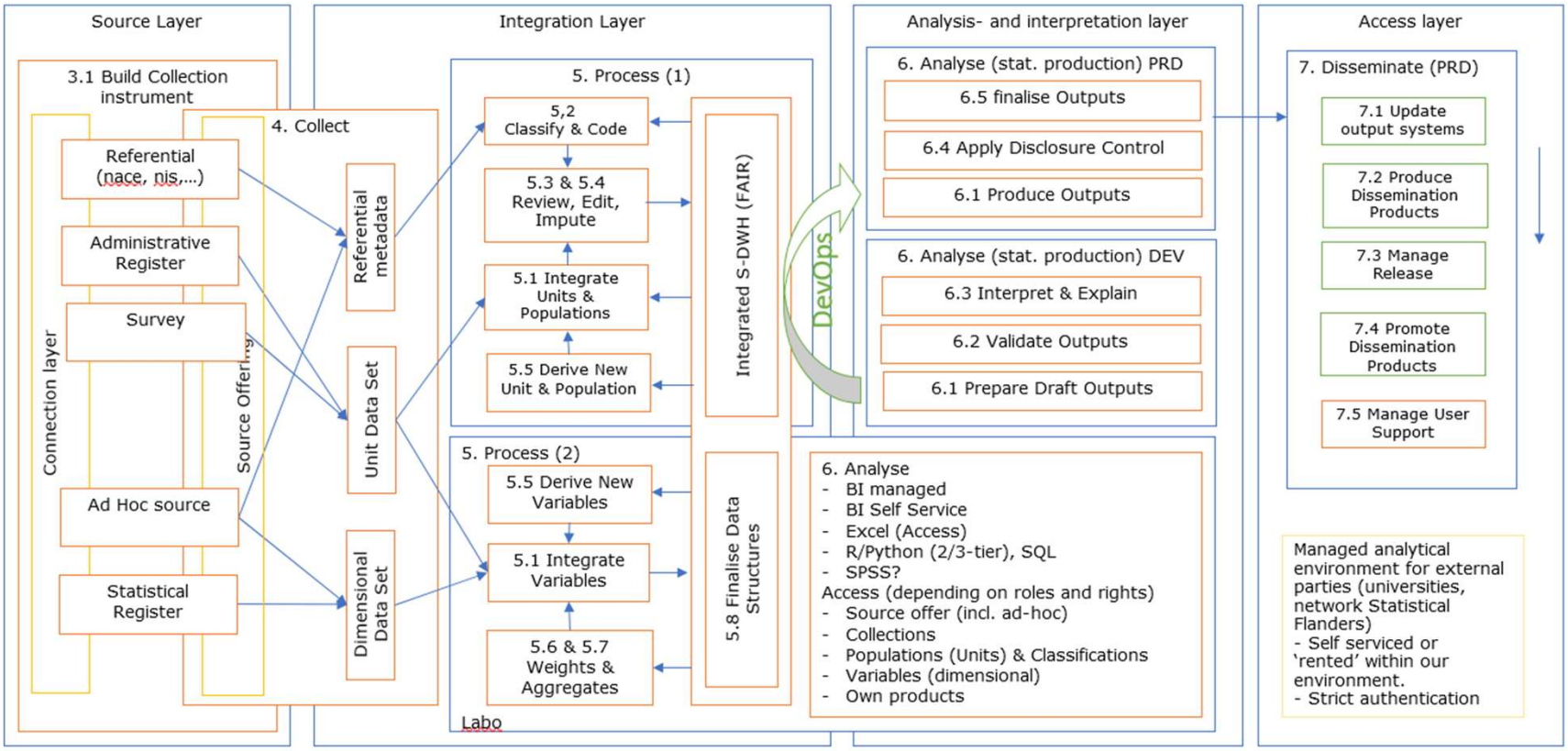
Business requirements: GSBPM (processes)



STATISTIEK VLAANDEREN

Vlaamse overheid

GSBPM and the Layered CoE model

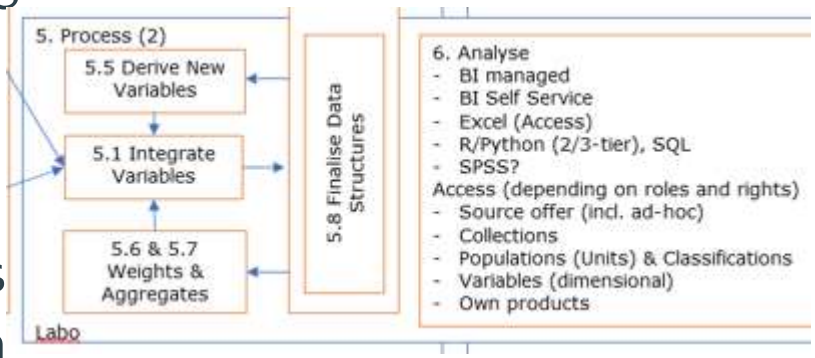


STATISTIEK VLAANDEREN

Vlaamse overheid

GSBPM and the CoE Layers

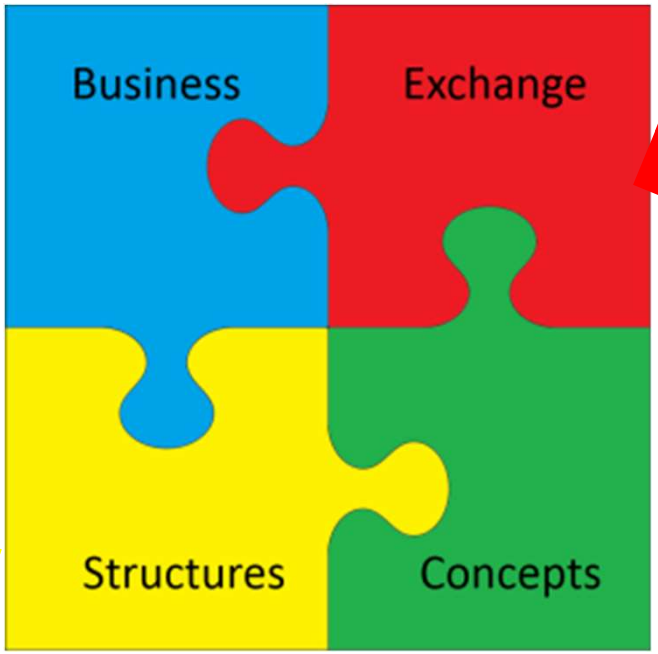
- Results:
  - 4 work phases of the GSBPM onto the 4 layers of CoE
    - 1 source connection layer for the collection instruments
  - Process phase
    - Management of *CategorySets, UnitTypes, Populations*
    - Integrating, transforming, deriving, analysing *Variables*
  - Laboratory environment
    - (partial) Process and Analysis phases
    - Emphasis on extreme flexibility with tooling (R, Python, SPSS, SAS,…)
  - Statistics production (with DevOps)



GSIM Groups

Capture designs & plans of statistical programs.

Identification of a Statistical Need.



Catalogue information that comes in and out the statistical organization.

Describe the collection and dissemination of information.

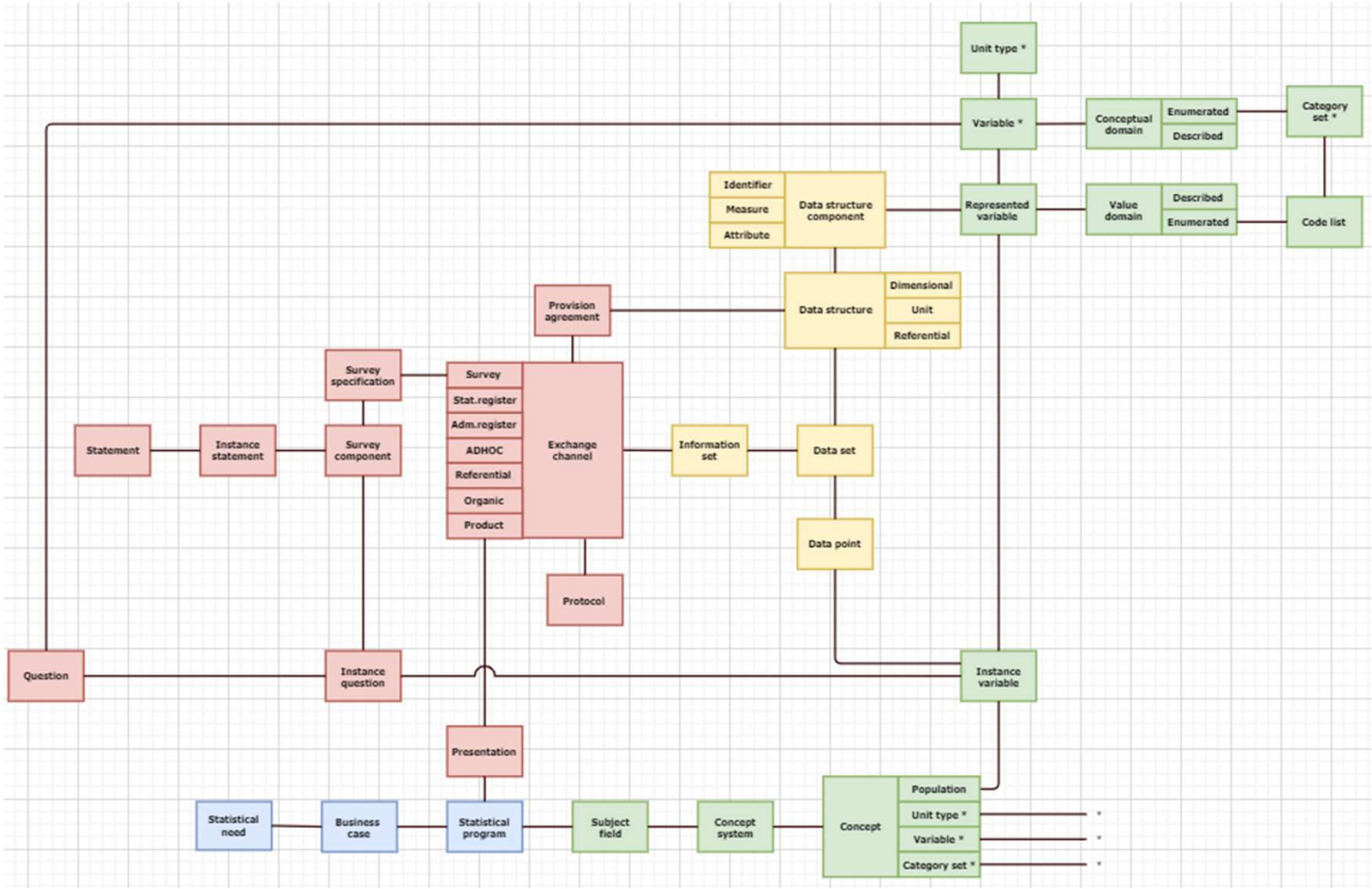
Describe and define the terms used in relation to information and its structure.

Define the meaning of data, providing an understanding of what the data are measuring.

GSIM Workshops

- GSIM has a very complete, detailed and large set of entities **(+-120)**
- Distilled to a **base model** after in-depth study **(+- 40)**
- Base model is discussed and reviewed by VSA
  - SME's
  - Data scientists
  - Data Strategist
  - S-DWH team
- This was done during 3 workshops

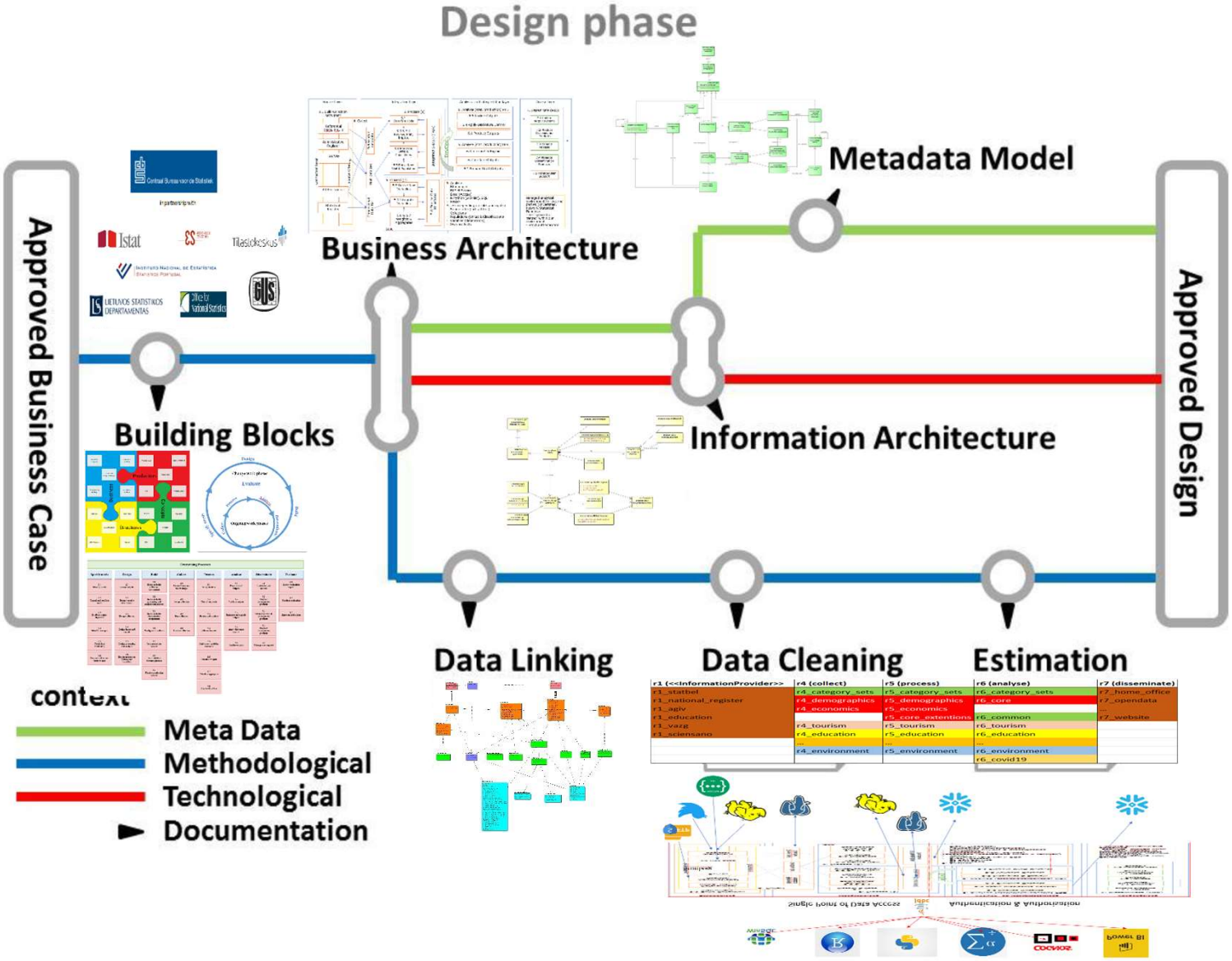
GSIM model @ VSA



STATISTIEK VLAANDEREN

Vlaamse overheid

COE Roadmap - Design phase components



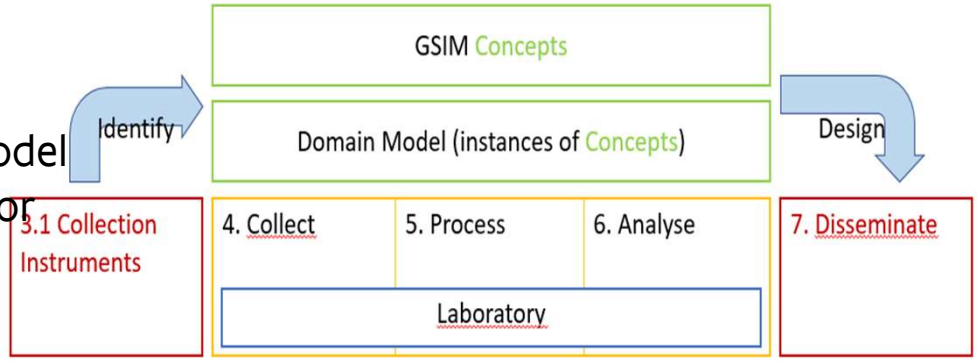
Practical phase: Matching the model with the reality

- Goal: organise data & processes around 2 core's:
  Persons and Organisations

- Organize workshops based on our incoming data files
  - exchange group : Why, when, from where, for what, outcome
  - concept & structure group : Data definitions, data format throughout the process

Integrating GSIM in GSBPM

- Concepts group:
  - GSIM level (M2)
    - Building blocks for Domain model
    - Search and navigation space for metadata.
  - Domain Model (M1)
    - Instances of Concepts
    - Navigation and access space for data.

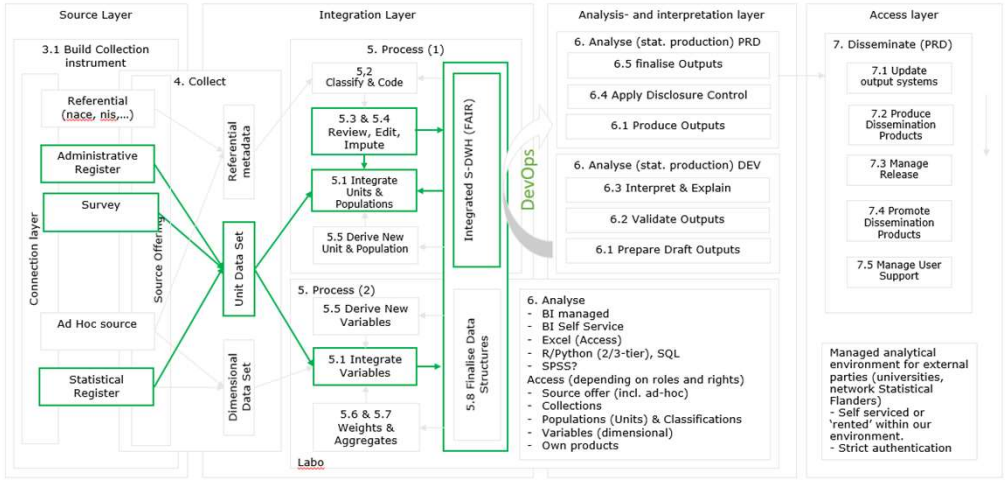


- Structures group:
  - Relates DataStructures to Represented Variables
  - Currently 3 types of structures (dimensional, unit, referential) coming from 4 sources: micro data; aggregated data; survey data; big data
  - Allows for grouping into InformationResources
    - Covering internal work phases (collect, process, analyse) and metdata (M2)
    - Further refinement according to SubjecField & ConceptGroup
    - Centralized authorization.
  - Independent of actual physical storage.
    - Questionable Data Structures
    - Service Level Agreement

- Exchange group
  - Manages input and output

Framework Templates – Microdata

- Units (microdata)
  - Collect phase
    - Identifying Variables
      - Identify key spaces
    - Populations
      - Variables
      - Identify key spaces
      - Time dependence
  - Process phase
    - Units (with mapping/matching structures where required/possible)
    - Populations
      - Basic time independent variables
      - Template structures for time dependent variables



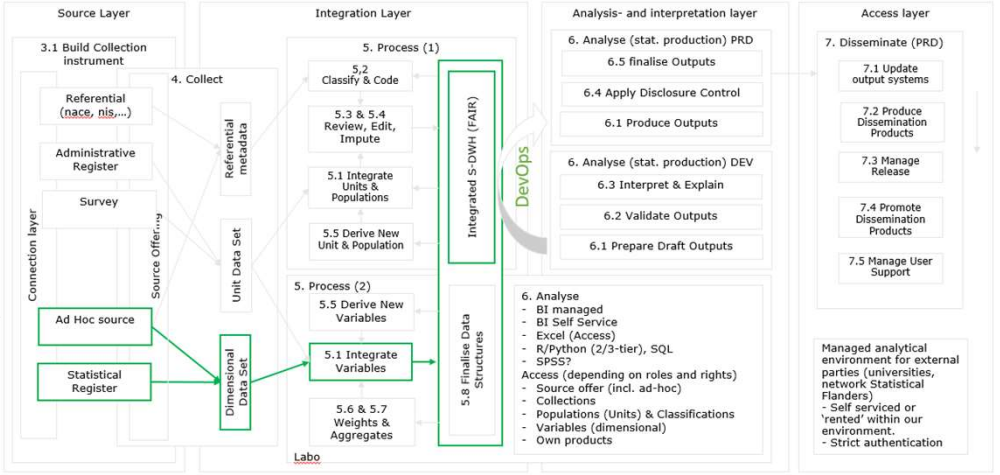
STATISTIEK VLAANDEREN

Vlaamse overheid

				Information Provider	Statbel			
				Exchange Channel	Jaarlijkse uitwisseling van demografische gegevens			
				Information Resource	Demografische datasets			
				Information Set	statbel_bevolking			
Unit Type	Variable	Variabele Omschrijving	Conceptual Domain	Tijdsaspecten? t0 geen historiek mogelijk ts Statistische waarnemingen te Event-stijl historiek	Population	Represented Variable	Instance Variable (voorlopig kolomnaam, suggesties welkom)	Value Domain
Persoon								
	vu_persoon	Identificatie van Persoon	Personen	t0	Bevolking	vu_persoon_rrnr_pseudo	ID_REC_C	Gepseudonymiseerd Rijksregisternummer
	vu_moeder	Identificatie van de Moeder	Personen	t0				
	vu_vader	Identificatie van de Vader	Personen	t0				
	vd_geboorte	Geboortedatum	Dag	t0	Bevolking	vd_geboorte	concat(geb_maand, geb_jaar, ge	Datum (yyyyMMdd')
	ve_geslacht	Geslacht	Geslacht	t0	Bevolking	ve_geslacht_statbel	geslacht_code	Geslacht-> sleutelruimte 1(statbel)
	ve_geboorteplaats	Geboorteplaats	Gemeentes - Subdomein van Bestuurlijke Indeling	t0				
	ve_geboorteland	Geboorteland ("moederland")	Land	t0				
	vd_overlijden	Sterfdatum	Dag	t0				
	ve_gemeente_overlijden	Gemeente van overlijden	Gemeentes - Subdomein van Bestuurlijke Indeling	t0				
	ve_hoofdverblijfplaats_overlijden	Hoofdverblijfplaats bij overlijden.	*Gemeentes* - Subdomein van Bestuurlijke Indeling	t0				
	vu_partner_overlijden	Identificatie van Partner bij overlijden	Personen	t0				
	ve_eerste_nationaliteit	Eerste Nationaliteit	Nationaliteit (of Land?)	t0	Bevolking	ve_eerste_nationaliteit_num3	fst_nat_code	Land-> sleutelruimte 1(num3)
	vd_eerste_nationaliteit	Datum eerste nationaliteit	Dag	t0	Bevolking	vd_eerste_nationaliteit	concat(fst_nat_jaar,fst_nat_maar	Datum (yyyyMMdd')
	ve_nationaliteit_overlijden	Nationaliteit	Nationaliteit (of Land?)	t0				
	ve_burgstat_overlijden	Burgerlijke Staat bij overlijden	Burgerlijke Staat	t0				
	vd_waarneming	Datum waarneming voor tijdsafhankelijke variabelen	Dag	ts	Bevolking	vd_waarneming	<jaar>'-12'31'	Datum yyyy'-12-31
	vu_huishouden	Huishouden	Huishoudens	ts	Bevolking	vu_huishouden	id_huish	Huishouden id's bepaald door statbel
	ve_nationaliteit	Nationaliteit	Nationaliteit (of Land?)	ts	Bevolking	ve_nationaliteit_num3	nat_code nat_code_vj	Land-> sleutelruimte 1(num3)
	ve_hoofdverblijfplaats	Hoofdverblijfplaats	*Gemeentes* - Subdomein van Bestuurlijke Indeling - Subdomein van Adressen *Postcodes*	ts	Bevolking	ve_hoofdverblijfplaats_nis4 ve_hoofdverblijfplaats_nis4 ve_hoofdverblijfplaats_postcode	gem_code gem_code_vj post_code	Bestuurlijke Indeling-> sleutelruimte 1(NIS)-> level 4 (gemeentes) Postcodes
	ve_verwantschap	Verwantschap ten opzichte van referentiepersoon van het Huishouden	Verwantschappen	ts	Bevolking	ve_verwantschap(_statbel)	verwant_code	Verwantschap-> sleutelruimte 1 (verwantschapscodes volgens Statbel)

Framework Templates – Dimensional data

- Dimensional Data
  - Collect phase
    - Identify Variables Iden
      - Primary Key (if exist
    - Identify Variables
      - Time dependence
      - Aggregation candidate
  - Process phase (1)
    - Match (and expand) CategorySets & Codelists (see referential metadata)
  - Process phase (2)
    - Link (CategorySets) and Integrate (Metadata)



MVP : S-DWH Approach

- Build, Buy, Meet in the middle?
  - Build from scratch?
    - Complex and elaborate GSIM & GSBPM.
    - Limited in-house resources (duration)
    - Waterfall effect (esp. with required flexibility)
  - Buy?
    - Comm. Products not aligned with GSIM/GSBPM (MOF, CWM)
    - Limited flexibility
    - Steep learning curve
  - Meet in the middle?
    - Technologies & Techniques from classical DWH? (useable?)
    - What functional components do we actually need?

Wrap up

- It is good to combine the different methodologies from the beginning
  - Stimulate the interaction
  - Evaluate the limits of each methodology
- Use the methodology as a model
  - Don’t be afraid of making a simplified operational model of your ideal theoretical model

You can reach us at

- Marc.Callens@Vlaanderen.be
- Freddy.Maetens@Vlaanderen.be