European
Commission

GSIM overview

Mauro Scanu
ISTAT

ESTP Training Course

“Information standards and technologies for describing, exchanging and disseminating data and metadata”

Rome, 19-22 June 2018

Summary

- Introduction to GSIM
- Business part
- Concept part
    - Population and unit
    - Variable, represented variable and instance variable
    - Classifications and code lists
- Structure part: micro and macro data sets
- Exchange part

Generic Statistical Information Model

http://www1.unece.org/stat/platform/display/gsim/Generic+Statistical+Information+Model

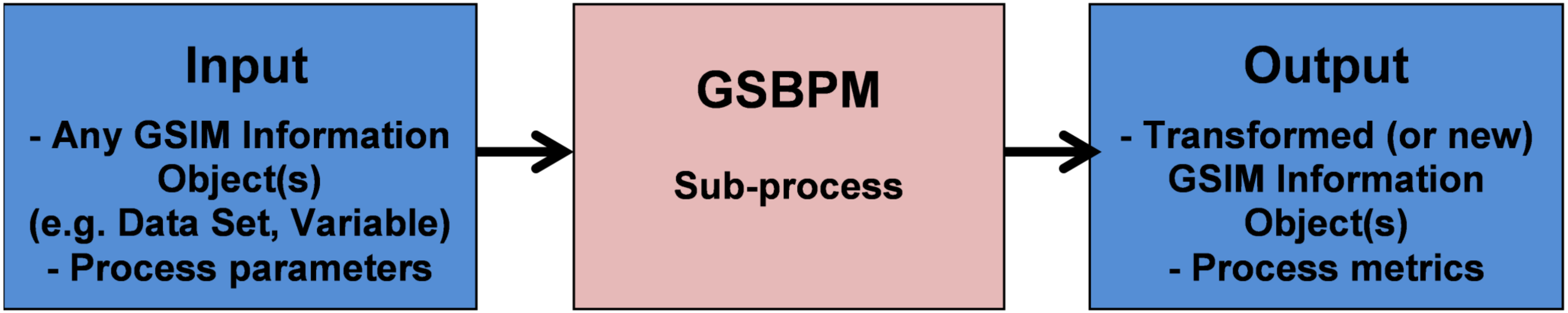
GSIM is the first internationally endorsed reference framework for statistical information. This overarching conceptual framework will play an important part in modernizing, streamlining and aligning the standards and production associated with official statistics at both national and international levels.

- **A reference framework for information objects (i.e. objects which specify information about the real world)**
- **Sets out definitions, attributes and relationships regarding information objects**

Relationship GSIM-GSBPM

GSIM provides the information object framework supporting all statistical production processes such as those described in the Generic Statistical Business Process Model (GSBPM)

GSIM and GSBPM are complementary



Input

- Any GSIM Information Object(s) (e.g. Data Set, Variable)
- Process parameters

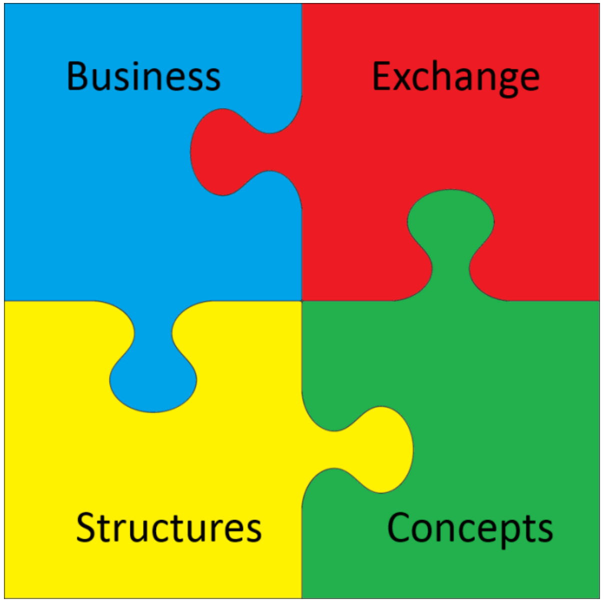
GSBPM

Sub-process

Output

- Transformed (or new) GSIM Information Object(s)
- Process metrics

GSIM



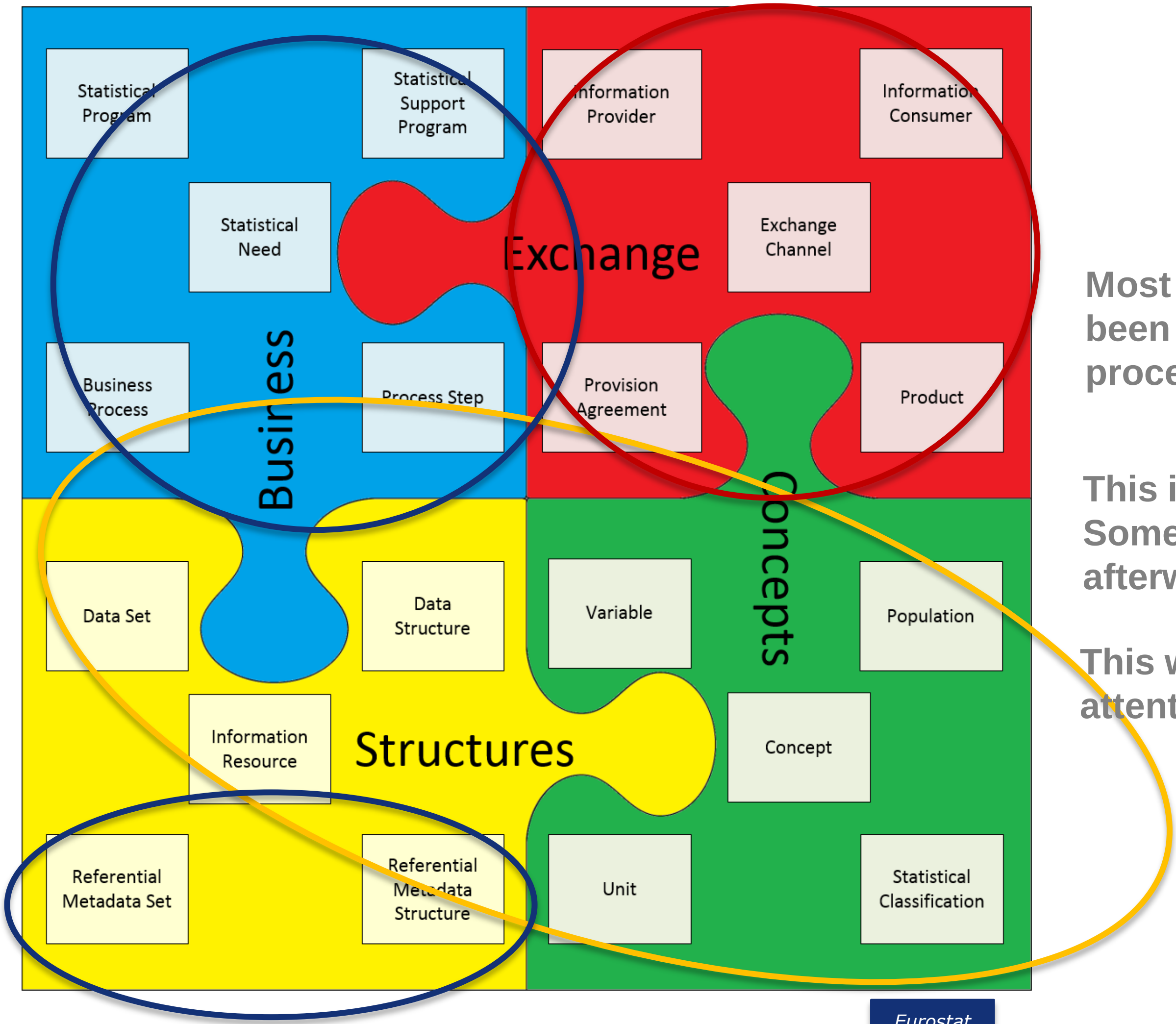
The Business group is used to capture the **designs and plans of statistical programs**, and the processes undertaken to deliver those programs. This includes the identification of a **Statistical Need**, the **Business Processes** that comprise the **Statistical Program** and the evaluations of them.

The Exchange group is used to catalogue the **information that comes in and out** of a statistical organization via Exchange Channels. It includes objects that describe **the collection and dissemination** of information.

The Concepts group is used to define the **meaning of data**, providing an understanding of what the data are measuring.

The Structures group is used to **describe and define the terms** used in relation to information and its structure.

Again on GSIM and GSBPM

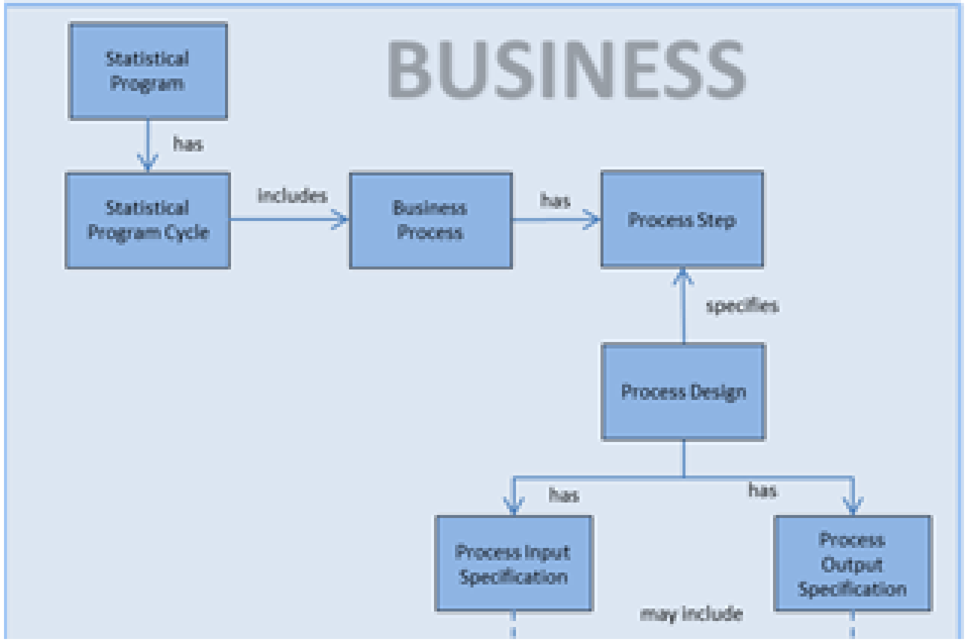


Most concepts in this part have already been detailed in GSBPM. A focus on process steps will be given later

This is also mostly covered in GSBPM. Some elements will be described afterwards

This will be the part where most of the attention will be posed

GSIM and Business



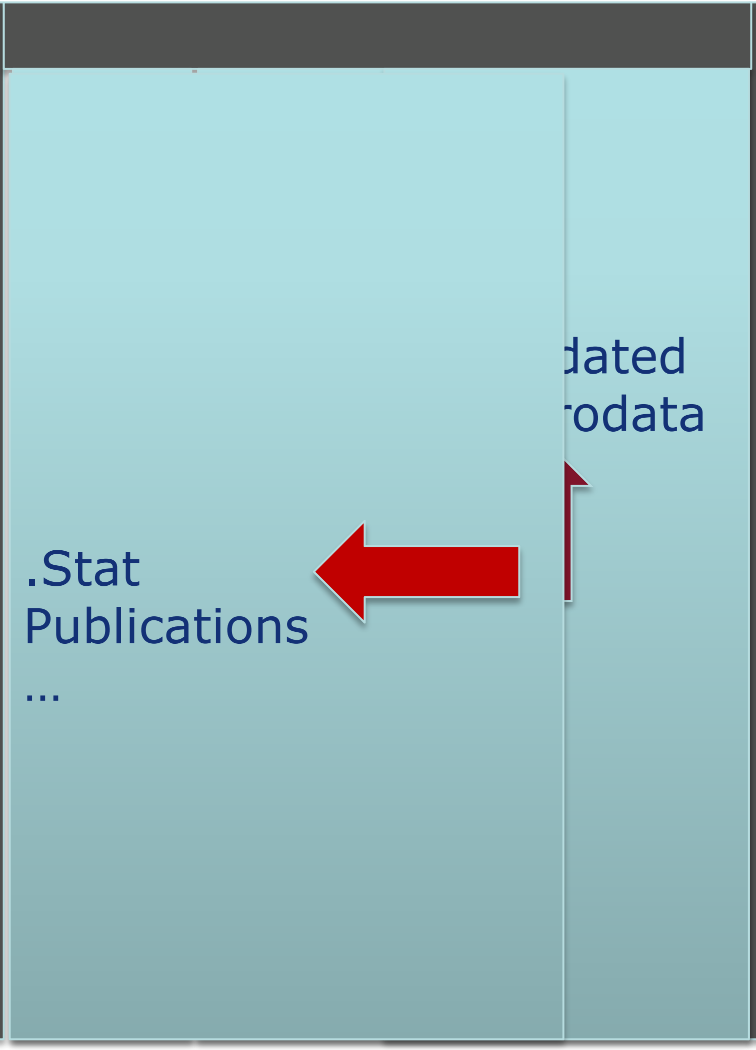
The level of detail in GSIM is different with respect to the level of detail in GSBPM. GSBPM details the subphases of each process.
GSIM pays more attention to the relationship between concepts, as well as the Process Input and Output Specifications

A detailed comparison between GSIM and GSBPM concepts is already under analysis

A statistical organization initiates a *Statistical Program*. The *Statistical Program* corresponds to an ongoing activity such as a survey or an output series and has a *Statistical Program Cycle* (for example it repeats quarterly or annually).

The *Statistical Program Cycle* will include a set of *Business Processes*. The *Business Processes* consist of a number of *Process Steps* which are specified by a *Process Design*. These *Process Designs* have *Process Input Specifications* and *Process Output Specifications*.

GSIM for structural metadata: where and what

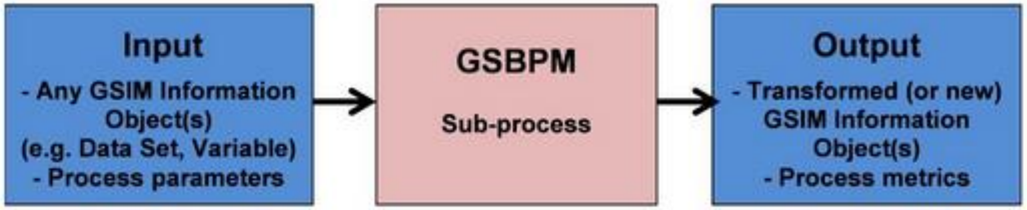


dated
rodata

.Stat
Publications
…

The GSIM part on structural metadata

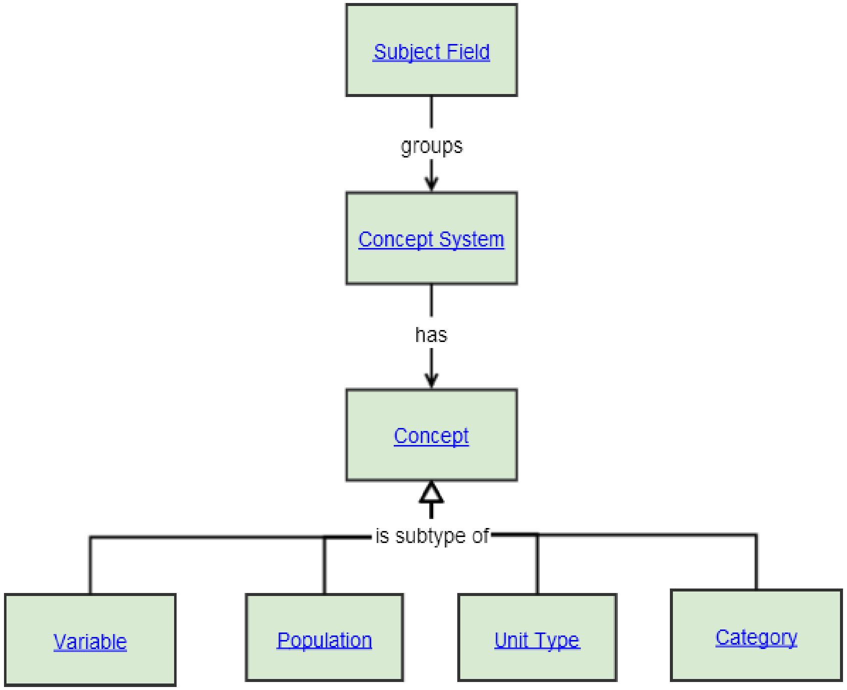
In order to give a proper definition to data inside a data structure GSIM is very convenient and effective because it describes data in their process flow


Input
- Any GSIM Information Object(s)
(e.g. Data Set, Variable)
- Process parameters
GSBPM
Sub-process
Output
- Transformed (or new) GSIM Information Object(s)
- Process metrics

Hence, GSIM defines all the ingredients useful for defining in a sound statistical way a data output.

Let’s look at the most important items that help in describing a statistical data output that are present in GSIM, and let’s give some comments

Foundational information (Concept)



60. *At an abstract level, a Concept is defined in GSIM as a 'unit of thought differentiated by characteristics'. Concepts are used in different ways throughout the statistical lifecycle, and each different role of a Concept is described using a different information object (which are subtypes of Concept).*

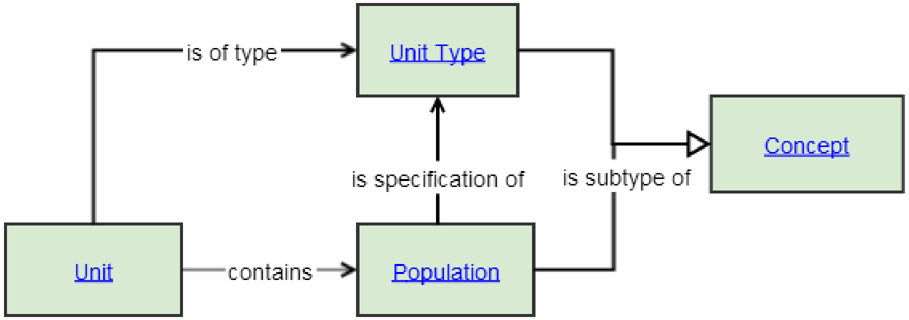
Comment: studying the variability of one characteristic (as well as the association of two or more characteristics) observed on the units of a population by means of aggregate output is the core of statistics

(a) As a characteristic (name of a statistical variable) observed on a population

(b) As a *Unit Type* or a *Population.*

(c) As a *Category* of a classification

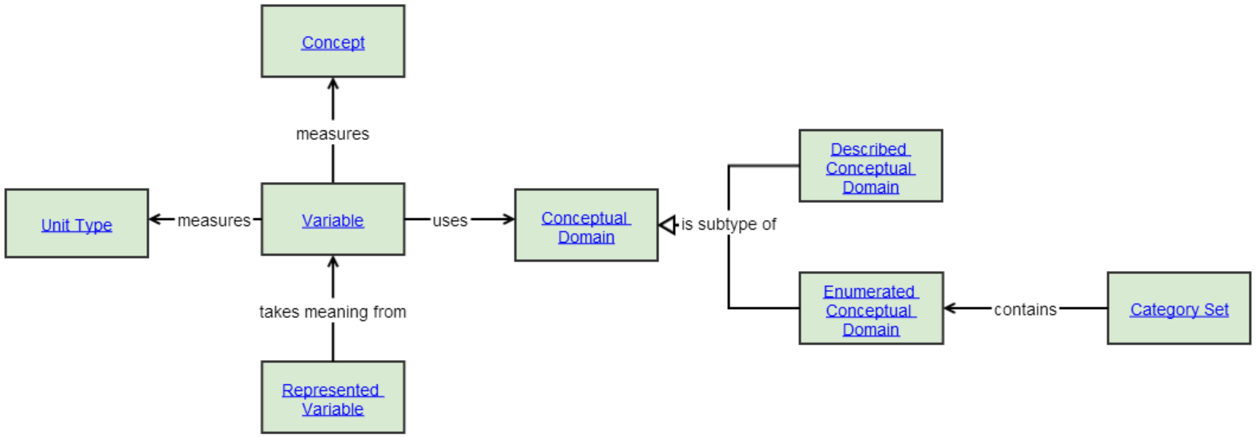
Population and unit



Comment: when dealing with a data set, we consider only one notion of population: the reference population of the data set

62. *There are several kinds of Populations depending on what Process Step it is used in. For example a statistical organization may refer to a target, survey, frame, or analysis population. The objects of interest in a statistical process are Units (for example, a particular person or a business). Data are collected about Units. There are two ways in which a unit is specified in the model. A Unit is an individual entity associated with a Population about which information may be obtained. A Unit Type (for example persons or businesses) is a way of identifying an abstract type of Unit that a Variable is measuring*

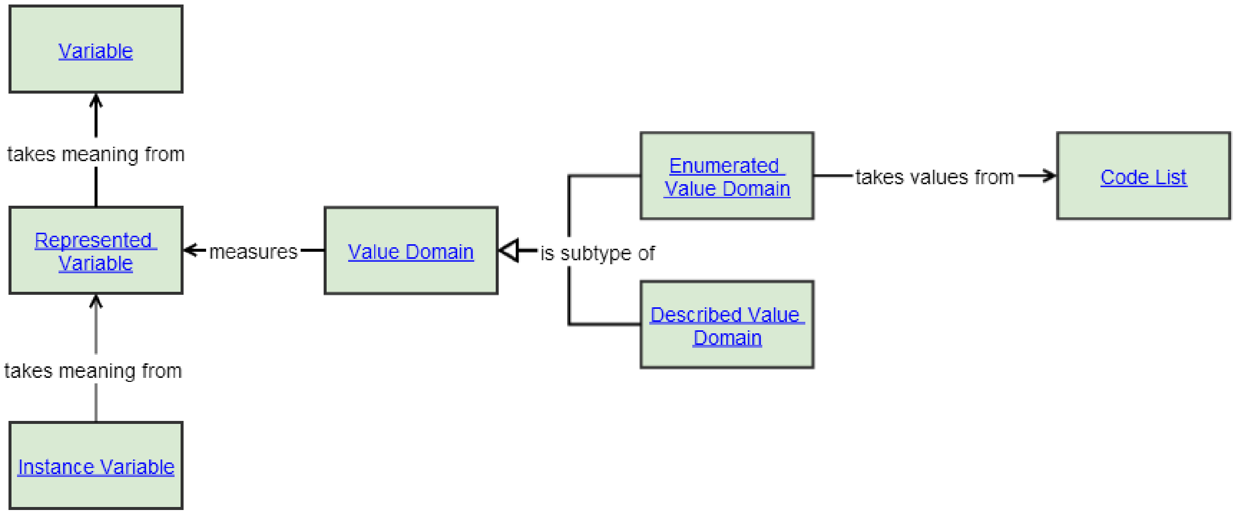
Variable



80. *When used as part of a Business Process, a Unit Type defining a Population is associated with a characteristic. The association of Unit Type and a Concept playing the role of a characteristic is called a Variable (see Figure 14). For example, if the Population is adults in Netherlands, then a relevant Variable might be the Concept educational attainment combined with the Unit Type person.*

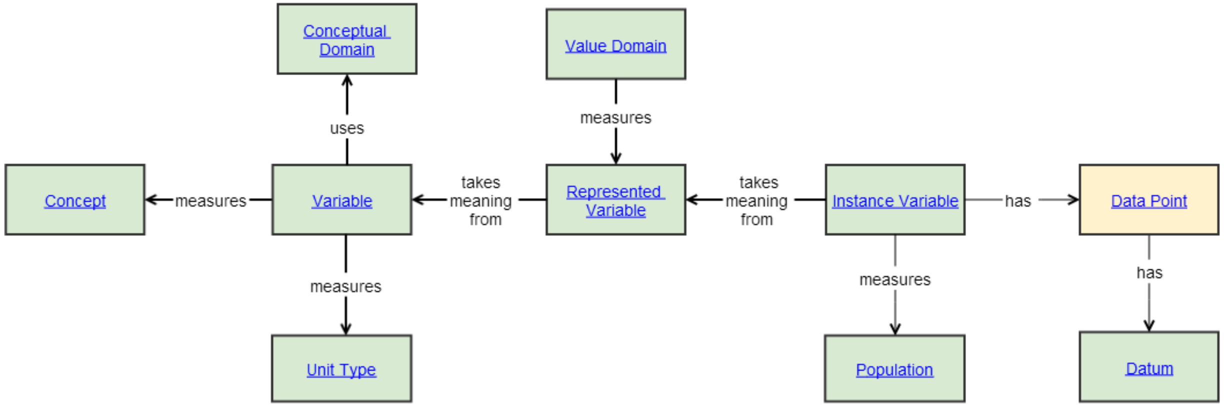
81. *The Variable (person's educational attainment) does not include any information on how the resulting value may be represented. This information (the Value Domain) is associated with the Represented Variable.*

Represented Variable



85. *The Represented Variable … adds information that describes how the resulting values may be represented through association with a Value Domain. While Conceptual Domains are associated with a Variable, Value Domains are associated with a Represented Variable. These two domains are distinguished because GSIM separates the semantic aspect (Conceptual Domain) and the representational aspect (Value Domain).*

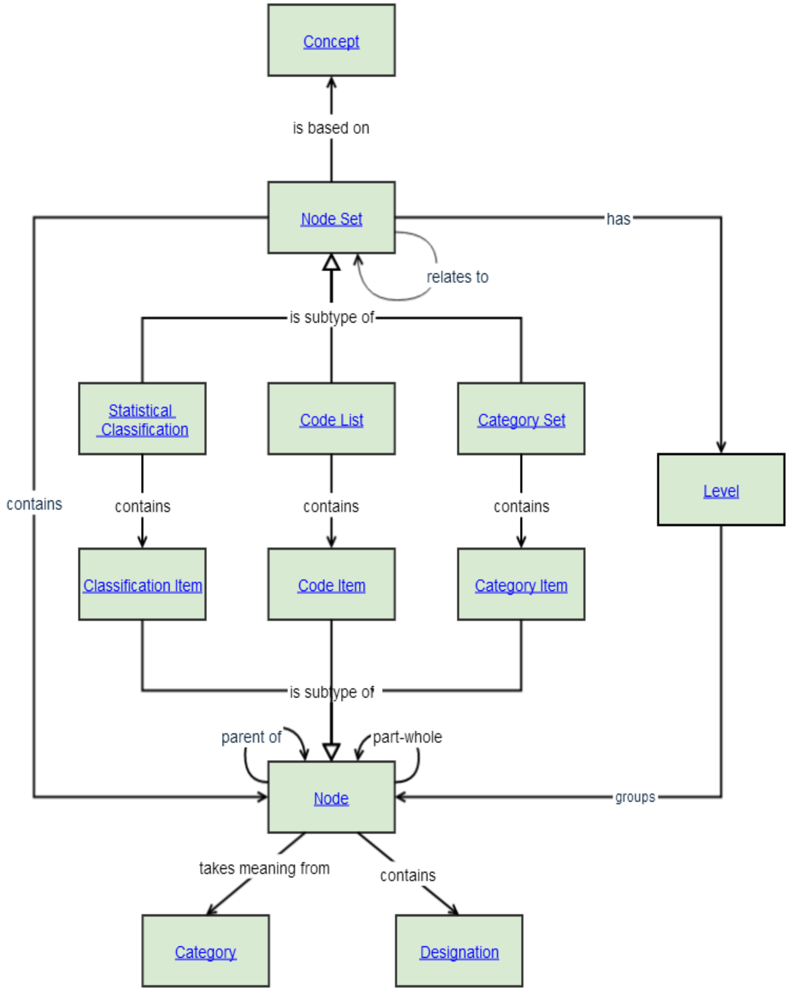
Instance Variable



88. *An Instance Variable (see Figure 16) is a Represented Variable that has been associated with a Data Set. This can correspond to a column of data in a database. For example, the "age of all the US presidents either now (if they are alive) or the age at their deaths" is a column of data described by an Instance Variable, which is a combination of the Represented Variable describing "Person's Age" and the Value Domain of "decimal natural numbers (in years)".*

89. *A Datum is contained within a Data Point in a Data Set. It may be defined by the measure of a Value Domain associated with a describing Instance Variable, combined with the link to a Unit (for unit data), or a Population (for dimensional data).*

Classifications, code lists and category sets

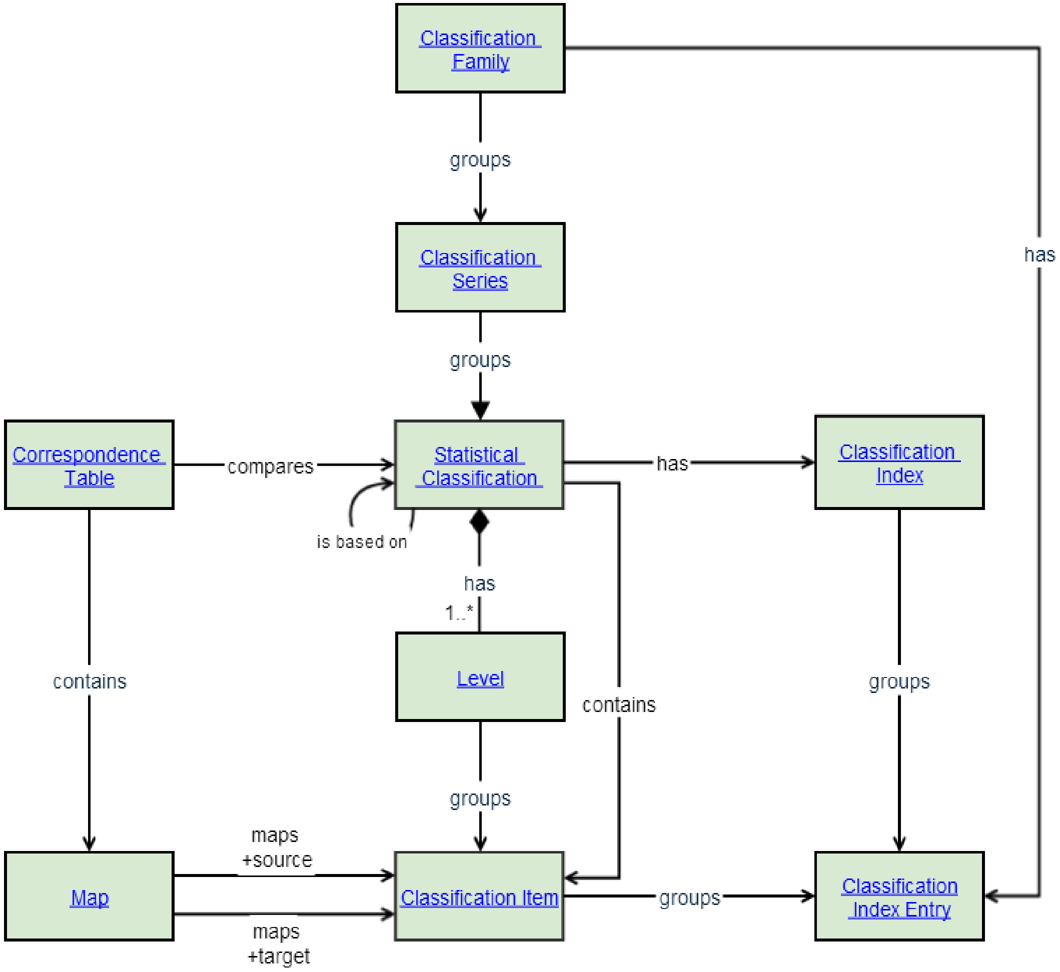


64. *A Category Set is a set of Category Items, which contain the meaning of a Category without any associated representations.*

65. *In a Code List, the Code Items contain the meaning of the Categories combined with a Code representation.*

66. *A Statistical Classification is similar to a Code List. It combines the meaning of the Category with a Code representation. However the content of a Statistical Classification must fulfil certain criteria and have a certain status. The Classification Items must be mutually exclusive and jointly exhaustive for the Level at which they exist at in the Statistical Classification.*

Statistical classifications



71. A Classification Family is a group of Classification Series related based on a common Concept (e.g. economic activity). A Classification Series is an ensemble of one or more Statistical Classifications that are based on the same Concept. The Statistical Classifications in a Classification Series are related to each other as versions or updates.

EXAMPLE:

Classification series: NACE

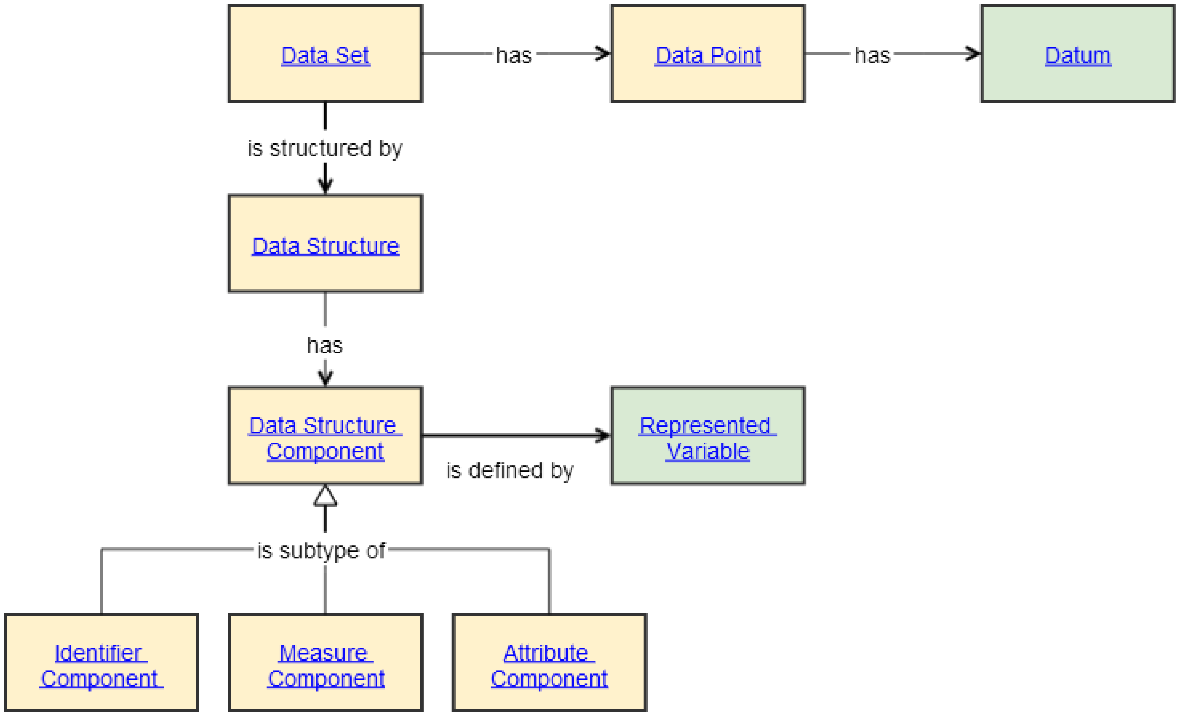
Statistical class. (version): NACE2007

Level: first digit

Class. item: A - Agriculture

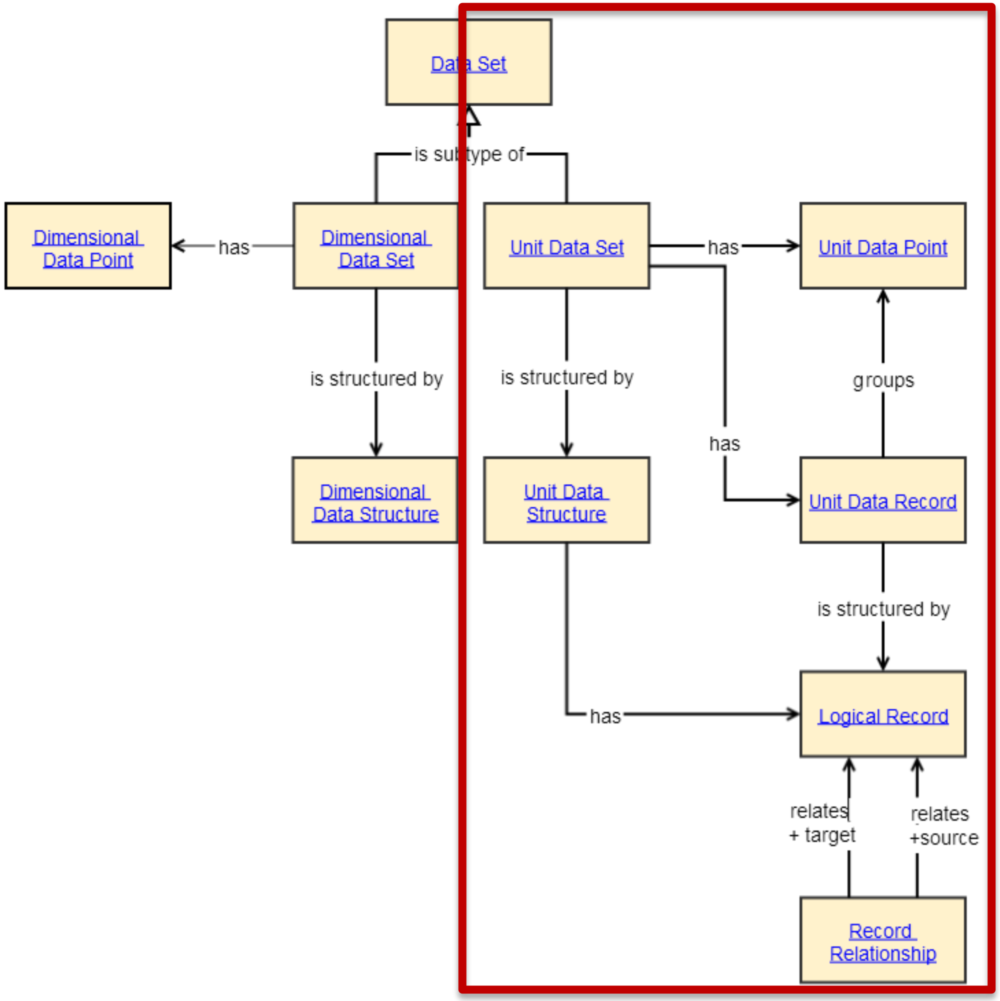
Data sets (Structures)

This diagram is valid for both micro and macro data sets



93. *A Data Set has Data Points. A Data Point is placeholder (for example, an empty cell in a table) in a Data Set for a Datum. The Datum is the value that populates that placeholder (for example, an item of factual information obtained by measurement or created by a production process). A Data Structure describes the structure of a Data Set by means of Data Structure Components (Identifier Components, Measure Components and Attribute Components). These are all Represented Variables with specific roles.*

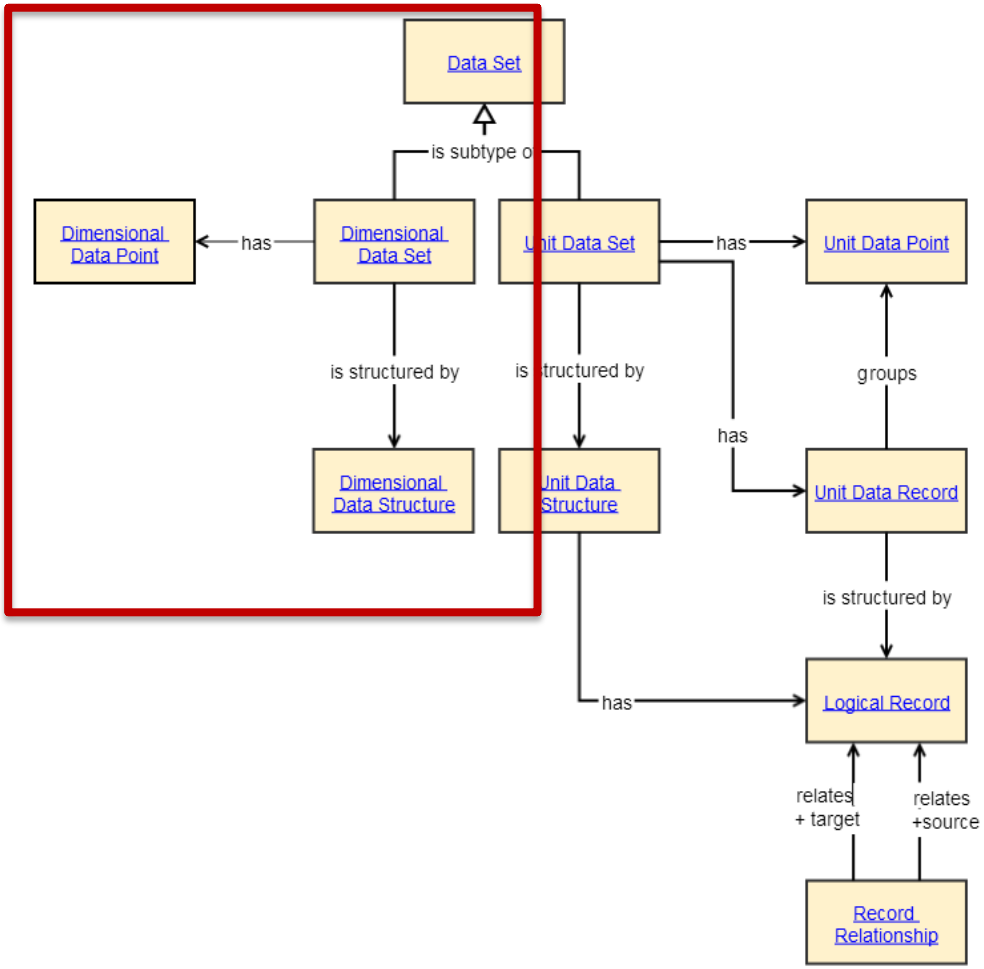
Dimensional and unit Data sets



102. *A Unit Data Structure describes the structure of a Unit Data Set by means of Represented Variables with specific roles. It distinguishes between the logical and physical structure of a Data Set. A Unit Data Set may contain data on more than one type of Unit, each represented by its own record type.*

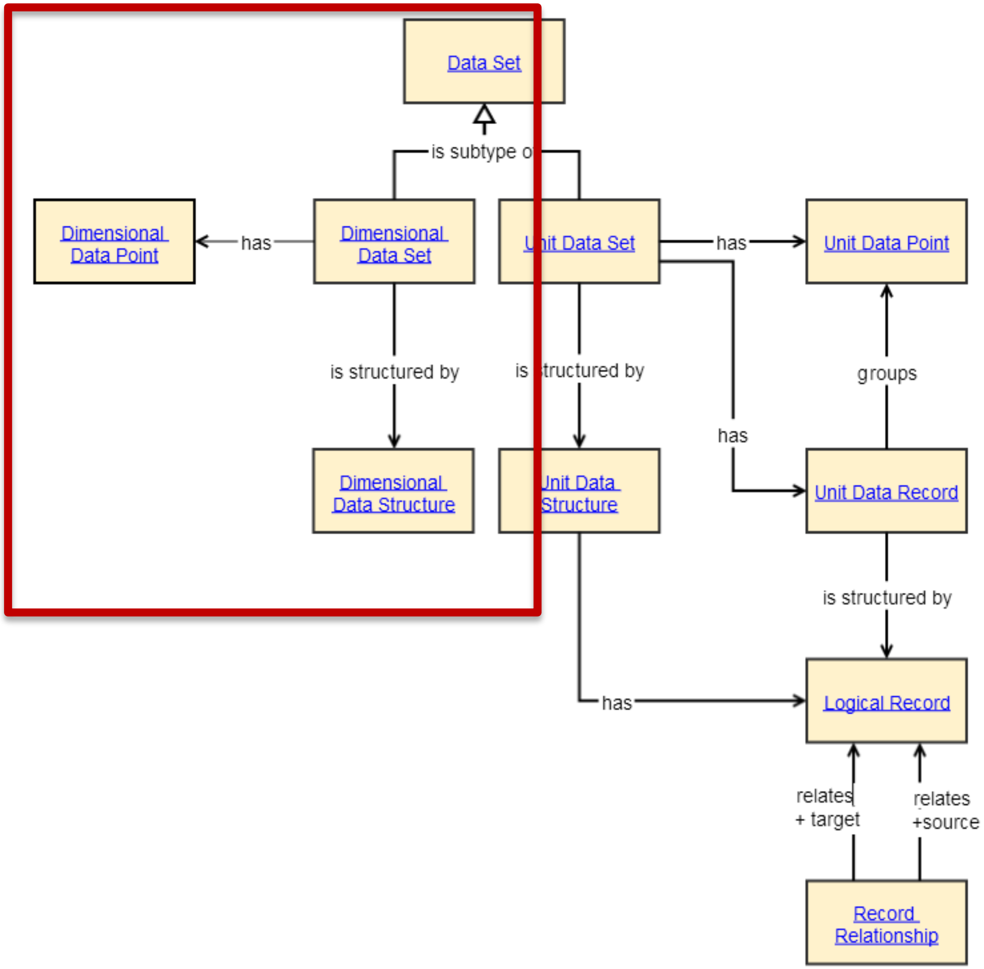
103. *Logical Records describe the structure of such record types, independent of physical features by referring to Represented Variables that may include a unit identification (for example, household number). A Record Relationship defines source-target relations between Logical Records.*

Dimensional and unit Data sets



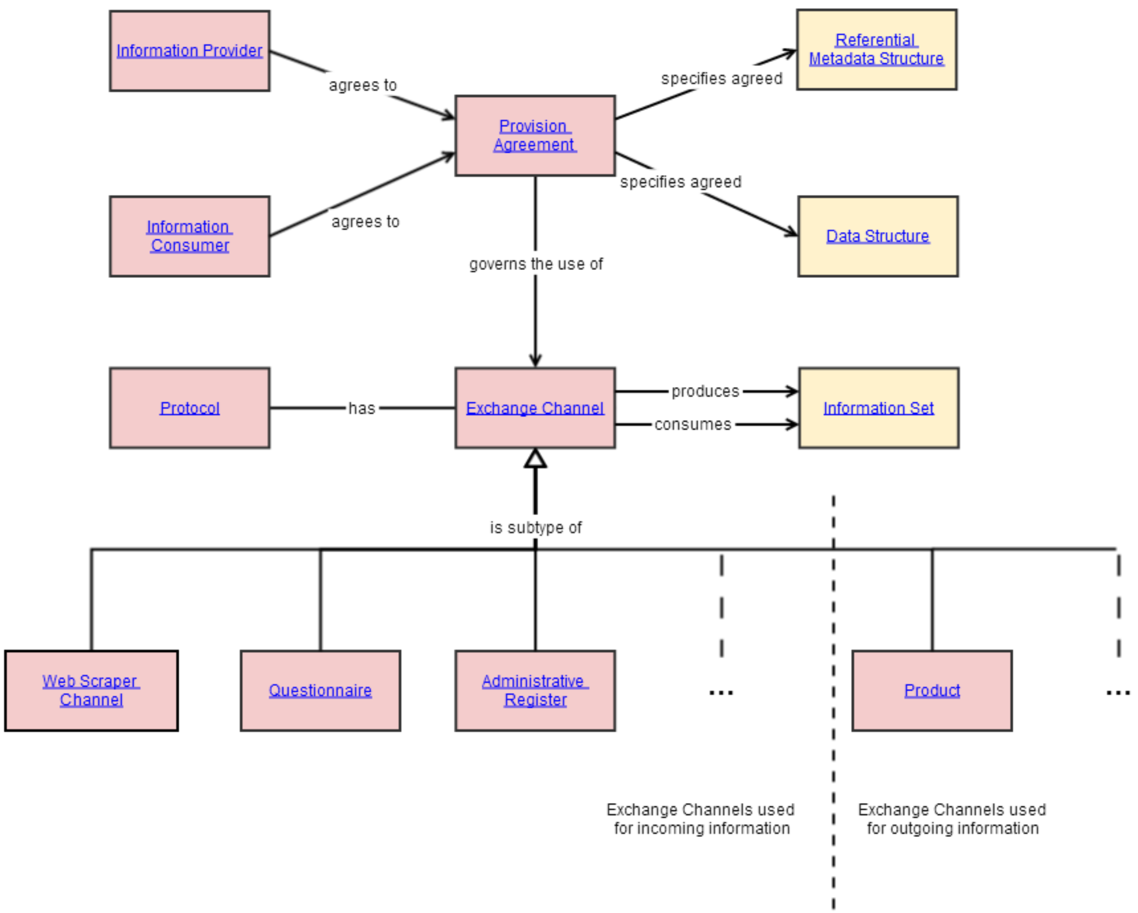
100. *The combination of dimensions contained in a Dimensional Data Structure creates a key or identifier of the measured values. For instance, country, indicator, measurement unit, frequency, and time dimensions together identify the cells in a cross-country time series with multiple indicators (for example, gross domestic product, gross domestic debt) measured in different units (for example, various currencies, percent changes) and at different frequencies (for example, annual, quarterly). The cells in such a multi-dimensional table contain the observation values.*

Dimensional and unit Data sets



101. *A measure is the variable that provides a container for these observation values. It takes its semantics from a subset of the dimensions of the Dimensional Data Structure. In the previous example, indicator and measurement unit can be considered as those semantics-providing dimensions, whereas frequency and time are the temporal dimensions and country the geographic dimension. An example for a measure in addition to the plain 'observation value' could be 'pre-break observation value' in the case of a time series. Dimensions typically refer to Represented Variables with coded Value Domains (Enumerated Value Domains), measures to Represented Variables with uncoded Value Domains (Described Value Domains).*

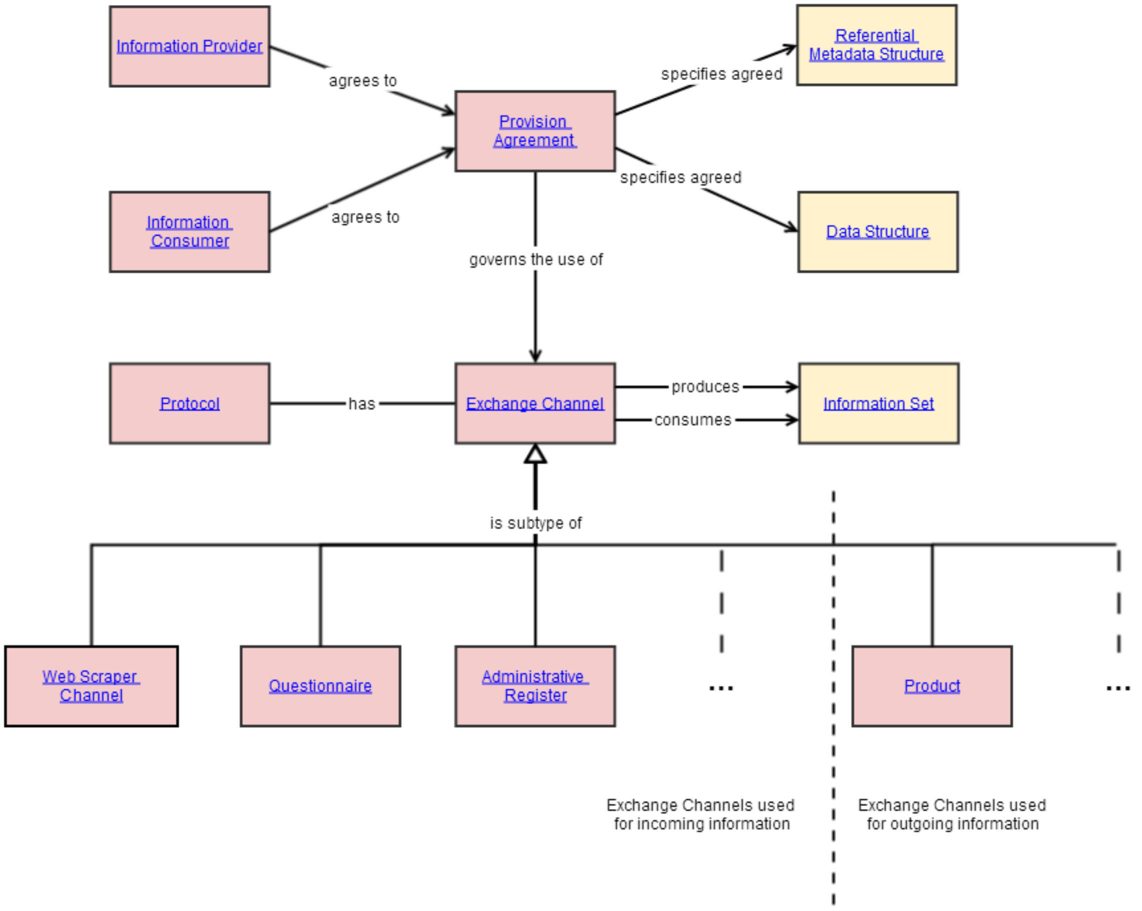
Exchange



39. *Statistics organizations collect data and referential metadata from Information Providers, such as survey respondents and providers of Administrative Registers, and disseminate data to Information Consumers, such as government agencies, businesses and members of the public. Each of these exchanges of data and referential metadata uses an Exchange Channel, which describes the means to receive (data collection) or send (dissemination) information…*

40. *Different Exchange Channels are used for collection and dissemination. Examples of collection Exchange Channels include Questionnaire, Web Scraper Channel and Administrative Register …*

Exchange



41. *The use of an Exchange Channel is governed by a Provision Agreement between the statistics office and the Information Provider (collection) or the Information Consumer (dissemination). The Provision Agreement, which may be explicitly or implicitly agreed, provides the legal or other basis by which the two parties agree to exchange data. The parties also use the Provision Agreement to agree the Data Structure and Referential Metadata Structure of the information to be exchanged.*

42. *The mechanism for exchanging information through an Exchange Channel is specified by a Protocol (e.g. SDMX web service, data file exchange, face to face interview).*