Metadata for statistical processes on registers: how to organize facts with GSIM

M. Scanu*, M. Scannapieco**, L. Tosco*, A.M. Bianco*, M. Riccio*
* Istat
**Agenzia per la cybersicurezza nazionale

Workshop on methodologies for official statistics - Roma, December 6, 2022

Overview

❑ Motivation of the work

❑ The 4 layers of Metadata along a statistical process in register-based statistics: from the register to the statistical result

❑ Representation of the concepts in ontologies

❑ Conclusions

Motivation of the work

Objective: The objective of this work is to define a layered metadata architecture integrated with the registers' setting that is serving three main purposes:

- ❑ Providing an explicit and well-defined statistical semantics to micro-data present in registers.
- ❑ Deriving structural metadata for different steps of the production process, up to the output statistical datasets of aggregate data organized in statistical data cubes that, in this case as well, have an explicit and well-defined statistical semantics.
- ❑ Governing the whole production pipeline from registers' data to output datasets by means of a well-defined and coherent metadata asset.

Effects:

- Verifiability
- Completeness of metadata
- Harmonization along a process and between processes (lineage and modularity)
- Guide for data users in organizing their output in terms of metadata (explainability and documentation)
- This is the base for supporting data users in organizing the aggregate data they desire

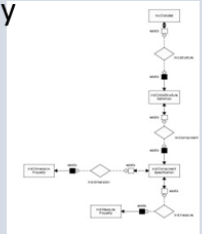
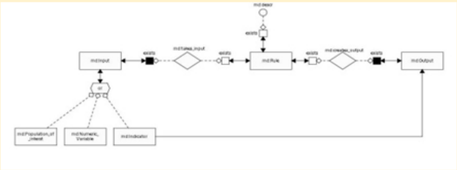
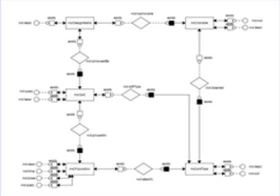
Tools used for getting the objective

A standard conceptual framework: The Generic Statistical Information Model is a conceptual model whose objective is the standard definition of the information objects that are input or output of the different steps of a production process. We used GSIM extensively, although we had the need to add some new concepts

Ontologies: In computer science, an Ontology is a conceptual specification of the domain of interest expressed through a formal language, shared and unambiguous

Main concepts used in this framework: Population, Unit, Variable, Data set, Data structure, Statistical program

The production line for register based statistics

	Intesional Level	Extensional Level	Data Level
4. Aggregate Data Set	Data Cube Ontology 	DataSet 1 Dimesion Property 1 Dimesion Property 2 Measure Property 1	
3. Macro Data	Ontology of macro data 	Population 1 Indicator 1 Variable 1	DW 1 DW N
2. Micro Data /Design Matrix	Ontology of micro data 	Population 1 Population 2 Unit 1 Unit Type 1 Variable 1	
1. Register Data	Domain Ontologies	Employee 1 Firm 1 Person 1 Family 1	Register 1 Register 2 Register N

Main characteristics of the ontologies in the 4 layers

1. Register data: all the concepts involved in the definition of the domain are formally inserted in an ontology. Concepts are not specialized according to a possible statistical role. A register is not a statistical data set. A register contains many possible micro data sets from which the statistical analysis can start

2. Micro data set/design matrix: In this step, there is the selection of concepts in the ontology according to their statistical role.

a) Start from the *concepts* in the register (*unit type*). E.g. Person, Household, Enterprise,… and select one of them

b) Specify the Universe: E.g. Resident households

c) Specify the Population (universe + time and area specifications): E.g. Resident households in Italy on the 1st of January 2020

d) List the Units in the population

e) Select the Variables of interest for all the units, i.e.

- The attributes of the chosen concept (e.g. household code)
- The attributes from connected concepts (e.g. area of residence, household overall income, number of children, main income source), sometimes as transformed variables

f) End up with a design matrix (units as rows, variables as columns)

Main characteristics of the ontologies in the 4 layers

3. Macro data (indicator): there are different kinds of aggregate data.

a) Synthesis of a statistical distribution
b) Comparison between aggregates

In the first case, the semantic of a synthetic value of a distribution is:

- Specify the Universe: E.g. Resident households
- Specify the Variable(s) (numeric) for which the synthesis is sought: e.g. net income
- Specify the aggregation method: e.g. Average

It was necessary to include, at this stage, additional concepts: e.g. the ontology needs that the variables to be analyzed are “joint”

In the second case, it is necessary to specify the characteristics of the comparison

- Ratios, densities,…
- Percent with respect to a previous year/quarter/…
- Index numbers (especially complex ones)

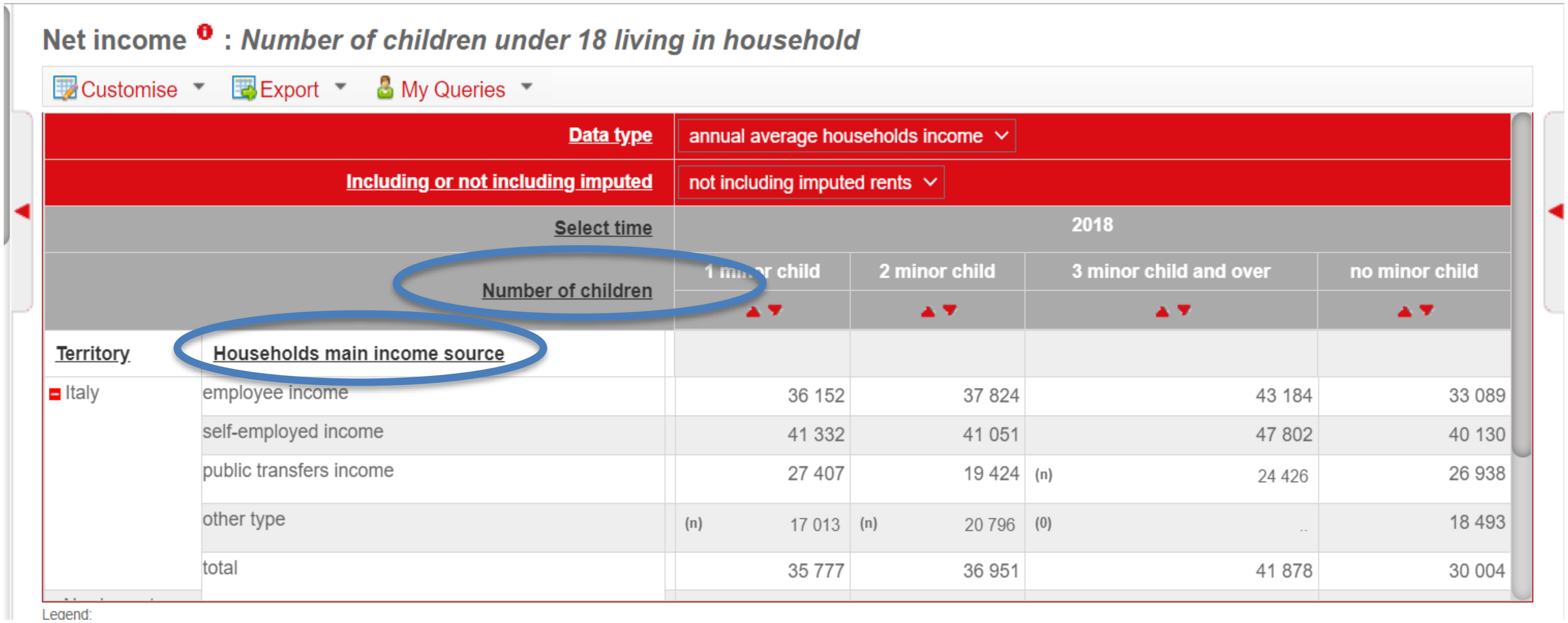
Main characteristics of the ontologies in the 4 layers

4. Macro data set:

a) Measure: the aggregate value with its semantics

b) Dimensions: the categorical variables that are used to find out partitions of the populations (including the geographical dimension), and a time dimension

- The universe in the measure *combined* with time and geographical dimensions make clear the population
- The other dimensions are usually the “conditioning variables”



Net income : *Number of children under 18 living in household*

Data type: annual average households income

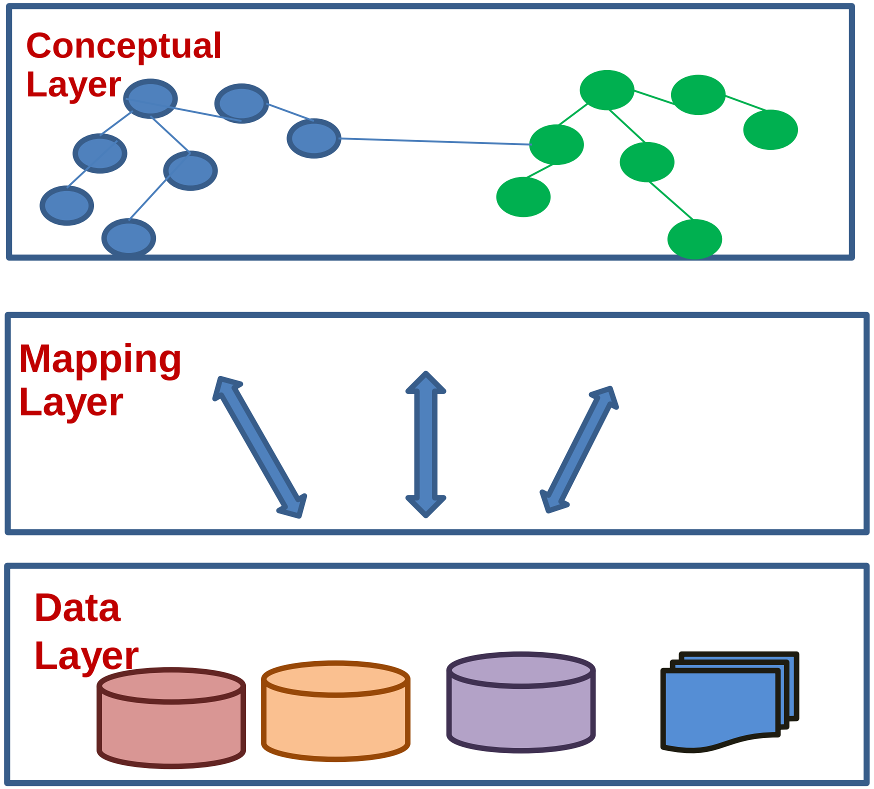
Including or not including imputed: not including imputed rents

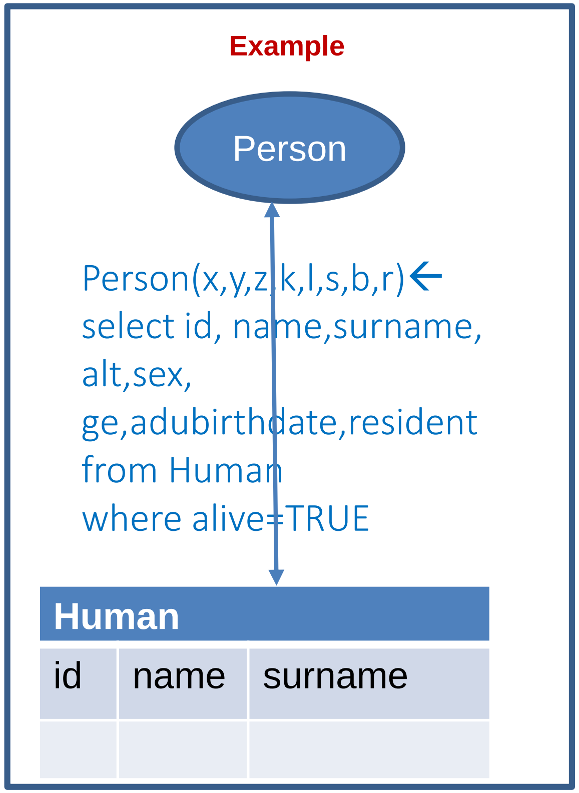
	Select time	2018			
	Number of children	1 minor child	2 minor child	3 minor child and over	no minor child
Territory	Households main income source				
Italy	employee income	36 152	37 824	43 184	33 089
	self-employed income	41 332	41 051	47 802	40 130
	public transfers income	27 407	19 424	(n) 24 426	26 938
	other type	(n) 17 013	(n) 20 796	(0) ..	18 493
	total	35 777	36 951	41 878	30 004

Legend:

O.B.D.A. paradigm

O.B.D.A. stands for Ontology Based Data Access. This paradigm allows to access data trought ontologies.





Person(x,y,z,k,l,s,b,r)←
select id, name,surname,
alt,sex,
ge,adubirthdate,resident
from Human
where alive=TRUE

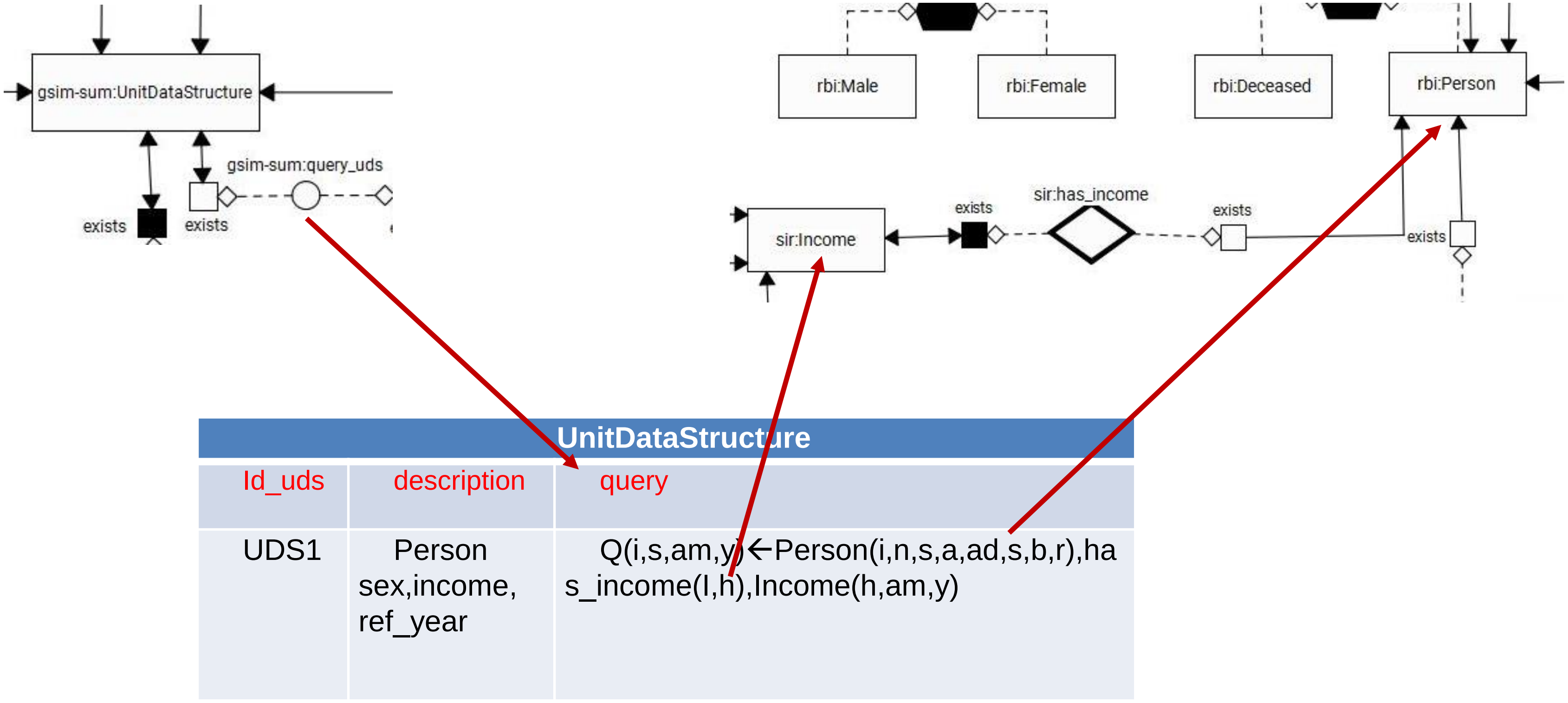
Human		
id	name	surname

In our case study, at the data layer, we found relational databases that stores both micro and macro meta-data that we represent as tables

Connection between ontologies of two layers

The 4 layers ontologies are connected, and these connections are established by queries.

E.g, consider the clear connection between the left ontology (on the design matrix, layer 2) and the right ontology (register ontology, layer 1) where the rows of the design matrix (persons among the resident persons) and its columns (income) are given



UnitDataStructure		
Id_uds	description	query
UDS1	Person sex,income, ref_year	Q(i,s,am,y)←Person(i,n,s,a,ad,s,b,r),has_income(I,h),Income(h,am,y)

Conclusions

- By the previous approach, there is just one conceptual framework for metadata relative to data produced in the different process stages. Harmonization along and between processes can refer to a single conceptual set up

- Some key concepts are organised as a «mixture» of other concepts with roles specified by the ontology. This is good for:
  - Checking completeness
  - Making harmonization easier
  - Defining the important assets for data search on data set, aggregates,…

- Only when necessary, new concepts have been added to the already available ones in GSIM (e.g., the multivariate nature of variables)

- Again: all of this refers only to metadata and how they represent data transfomation along a data production process.

Thank you for your attention