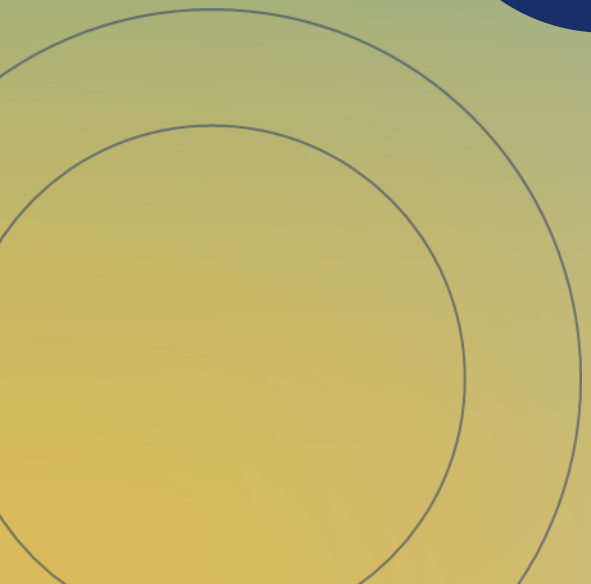
EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Semantic and ontologies of data sets along a data production process

Michele K. Riccio , Mauro Scanu

Italian National Institute of Statistics – Istat, Rome, Italy

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

The problem

How to describe statistical choices and requirements performed along a statistical process?

How to trace lineage of statistical data?

Our goals for description:

- Description at semantic level
- Not ambiguous
- Machine readable

❖ Is it possible to **retrieve data records** and to describe **semantic** of data, in the **same environment** ?

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Inside GSBPM

Preparation of data to be analyzed involves some phases of GSBMP.

We have to document statistical interpretations and choices from the beginning to the end of the process.

Anyway, data manipulations characterise mainly some phases, mainly those from the Collect to the Process phases

When our input data consist of a Statistical Register we have to consider that some manipulations have already been made:

- Integrate data
- Classify and code

Choices to prepare Registers have to be documented first.

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Data milestones

To analyze data from statistical point of view, we need to:

- Collect data or organize data in Statistical Registers
- Build a Design Matrix of microdata
- Compute the desired Macrodata: aggregated measures, indicators, derived variables

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Data milestones

Variable

Design matrix

Unit

Population

	X_1			X_k
u_1				
u_2				
		x_{ij}		
u_n				

measure

Aggregate data

Resident population on 1st January

Customise | Export | Draw chart | My Queries

Age	total		
Marital status	total		
Demographic data type	population on 1st January		
Select time	2017		
Gender	males	females	total
Territory			
Italy	29 445 741	31 143 704	60 589 445
Nord-ovest	7 832 094	8 271 788	16 103 882
Nord-est	5 664 202	5 972 900	11 637 102
Centro (I)	5 822 205	6 245 319	12 067 524
Sud	6 856 385	7 214 776	14 071 161
Isole	3 270 855	3 438 921	6 709 776

Resident population on 1st January

Source
Data source(s) used
Resident municipal population by age, sex and marital status
The English description of the source is not available at this time, for methodological details go to the Sigual system
Resident municipal population by age, sex and marital status

Data Characteristics
Other data characteristics
Data for previous years are available in the Inter censuses estimates theme
References to the territorial changes

How do these concepts interact?

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Description layers

Layer	describes	Represented by	Meaning
Input Data Layer	Data Collected or Statistical Registers	Domain Ontology	Semantic of domain to analyze
Design Matrix Layer	Design Matrix of microdata	Design Matrix meta Ontology	Statistical choices for microdata
Macrodata Layer	Macrodata: aggregated measures, indicators, derived variables	Aggregate Data meta Ontology	Statistical choices to aggregate microdata into macrodata

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Why meta ontologies ?

Our strong requirement: Description of statistical choices have to be **decoupled** from specific domains of input data.

In our approach meta ontologies are abstract and general, only their istances are specific for a single domain.

We could use also SparQL queries directly on Domain Ontology, but so statistical choices are coupled and embedded in one specific Data Domain.

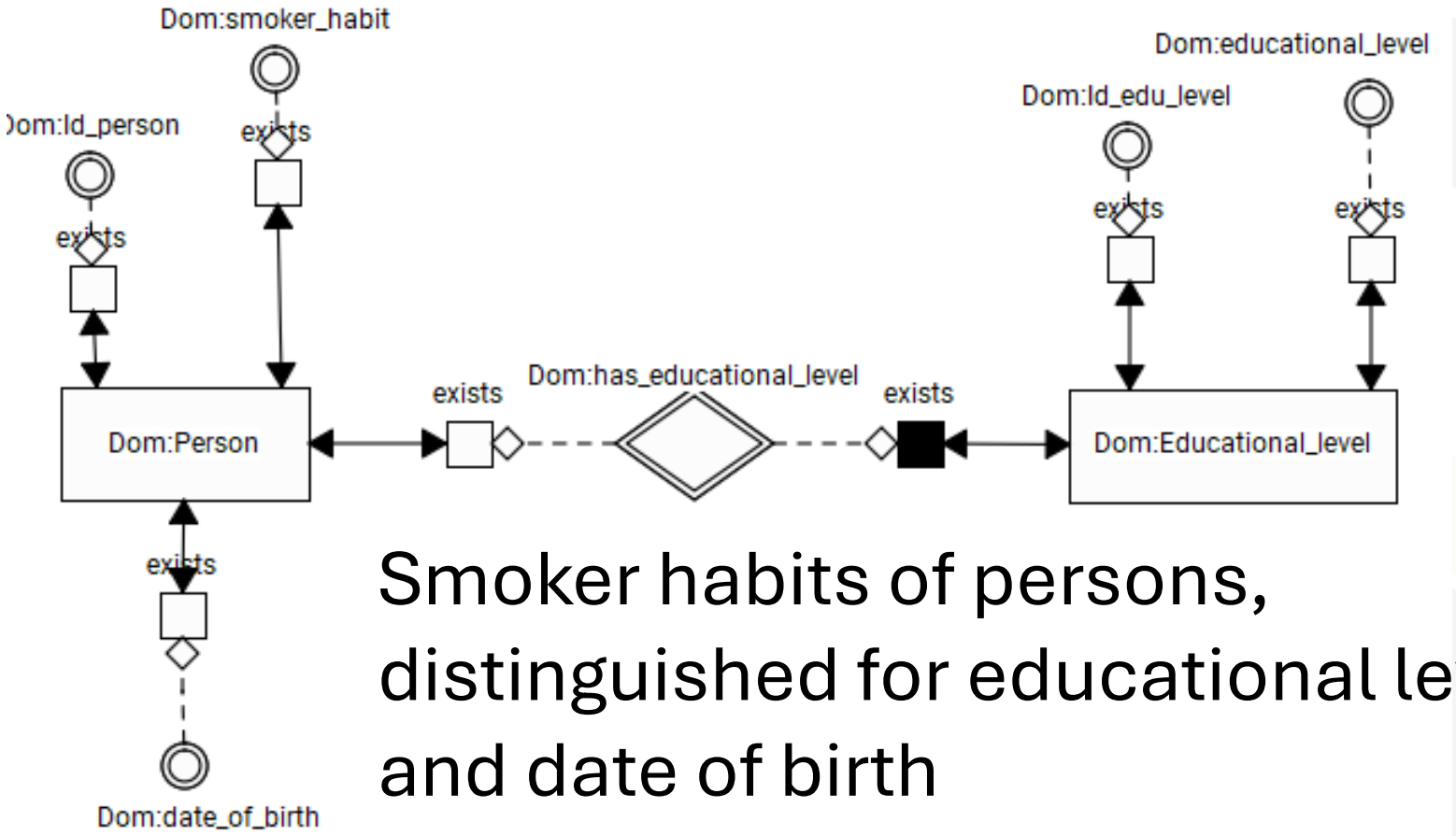
INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Example of Domain Ontology



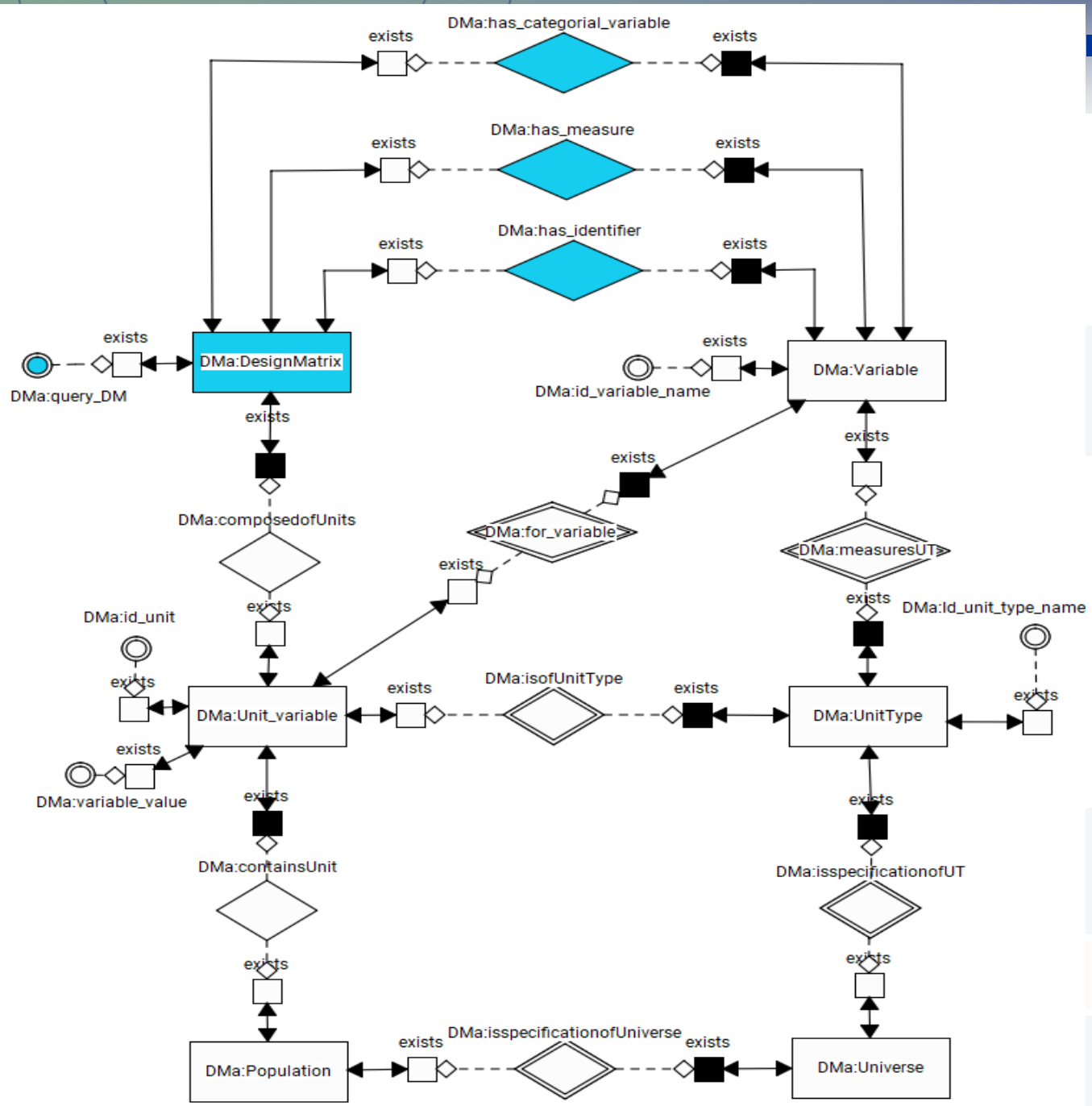
Smoker habits of persons, distinguished for educational level and date of birth

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

The conference is partly financed by the European Union

Design Matrix meta Ontology

Expressed in Graphol language



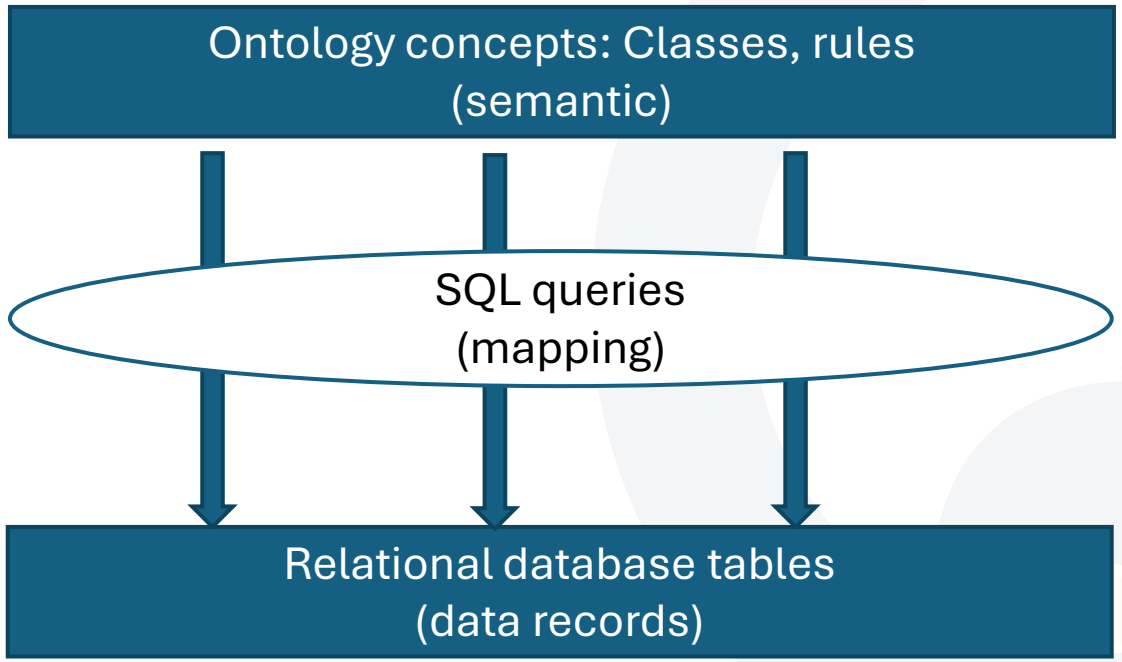
EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

OBDA approach

Ontology Based Data Access permits to:

Retrieve data records by ontologies

Statisticians can fill their matrices without accessing to databases.



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Ontology View

Extended ontology construct: It is a special class

Representing the concept of the group of records contained in that design matrix.

It consists of:

- A SparQL query containing statistical choices
- Each attribute of View corresponds to one output field of the query
- Any instance of view represents the set of records contained in one Design Matrix

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Conclusions

By this approach statistical choices are:

- Described in a formal way (SparQL)
- Usable by reasoning services
- Accessible machine-to-machine.

It is possible:

- ❖ to trace the **lineage** of **data** transformations and of statistical **logic** applied in the process
- ❖ and **access data** in the meantime.

It needs **custom software** to enable these capabilities for users.

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union