A Knowledge Representation Approach for Modeling Aggregates: A case study at ISTAT

Domenico Lembo, Antonella Poggi – Sapienza University of Rome
Roberta Radini, Michele Riccio - ISTAT
Valerio Santarelli – OBDA Systems Srl

DEPARTMENT OF COMPUTER, CONTROL, AND MANAGEMENT ENGINEERING ANTONIO RUBERTI

SAPIENZA
UNIVERSITÀ DI ROMA

Case study at ISTAT: the INTERSTAT project

Problem

As national statistical institute, ISTAT collects **aggregate** data, also called **macro-data**, coming from different public bodies, each allowing separate multidimensional analysis

→ how to derive synthetic indicators to support decision-makers?

→ need to integrate such data in order to enable a unified cross-border and cross-domain **multidimensional** analysis over them

Context

The ISTAT Integrated System of Statistical Registers (ISSR)

- a unified conceptual point of access to socio-demographic, territorial, and institutional registers

→ **micro-data** have been integrated and made interoperable through **ontologies**

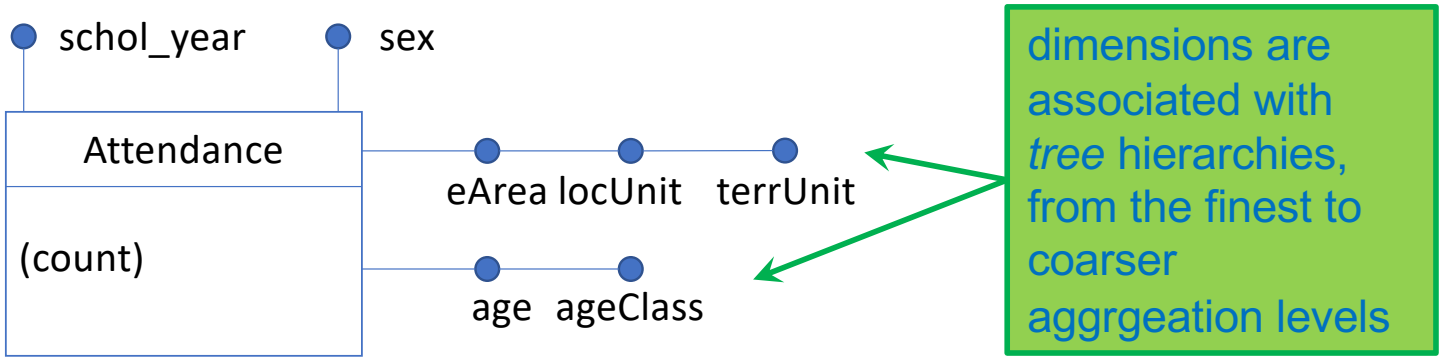
Motivating example

The INTERSTAT pilot “School for You” (S4Y):

- Goal: to define comparative indicators on the population of students by order of study, building upon various macro-data sets about school attendance in Italy and France
- Available data: number of students who attended a school in Italy since 2015, classified by
  - scholastic year
  - age groups (e.g., from 5 to 10 years old, from 11 to 14 years old, etc.),
  - sex
  - geographical location of schools → i.e., classified according to a standard mechanism, which associates:
    - schools to so-called enumeration areas, i.e., geographical areas used for censuses
    - enumeration areas to local administrative units
    - local units to territorial units at the third level of the EUROSTAT NUTS nomenclature1, denoted NUTS3 and corresponding, e.g., to Italian provinces and metropolitan cities or to French departments

State of the art: the DFM

In order to carry out the analysis though OLAP operators, we can model school attendance through a multidimensional cube, which we describe by means of the Dimensional Fact Model (DFM)



- Each fact (also called event) instantiating the cube represents the school attendance of a class of students characterized by a certain sholastic year, sex (male or female), age or class age and location, i.e., an enumeration area, a local unit, or a territorial unit

Limits of the DFM

- The schema does not say that the cube is referring only to students who attended a school in Italy since 2015 (such aspects are typically described in the documentation associated to the schema, often in an informal way)
- Important metadata end up only into the code of ETL procedures that extract source data and populate the data warehouse

→ how can one compare cubes? For instance, how can we know if it makes sense to compare the cube about the school attendance in Italy with a cube reporting «similar» data about French schools?

  – one should know from the representation model that both cubes contain data referring to the same period, i.e., since 2015, and described at the same level of granularity!

Proposal to overcome the limits of DFM

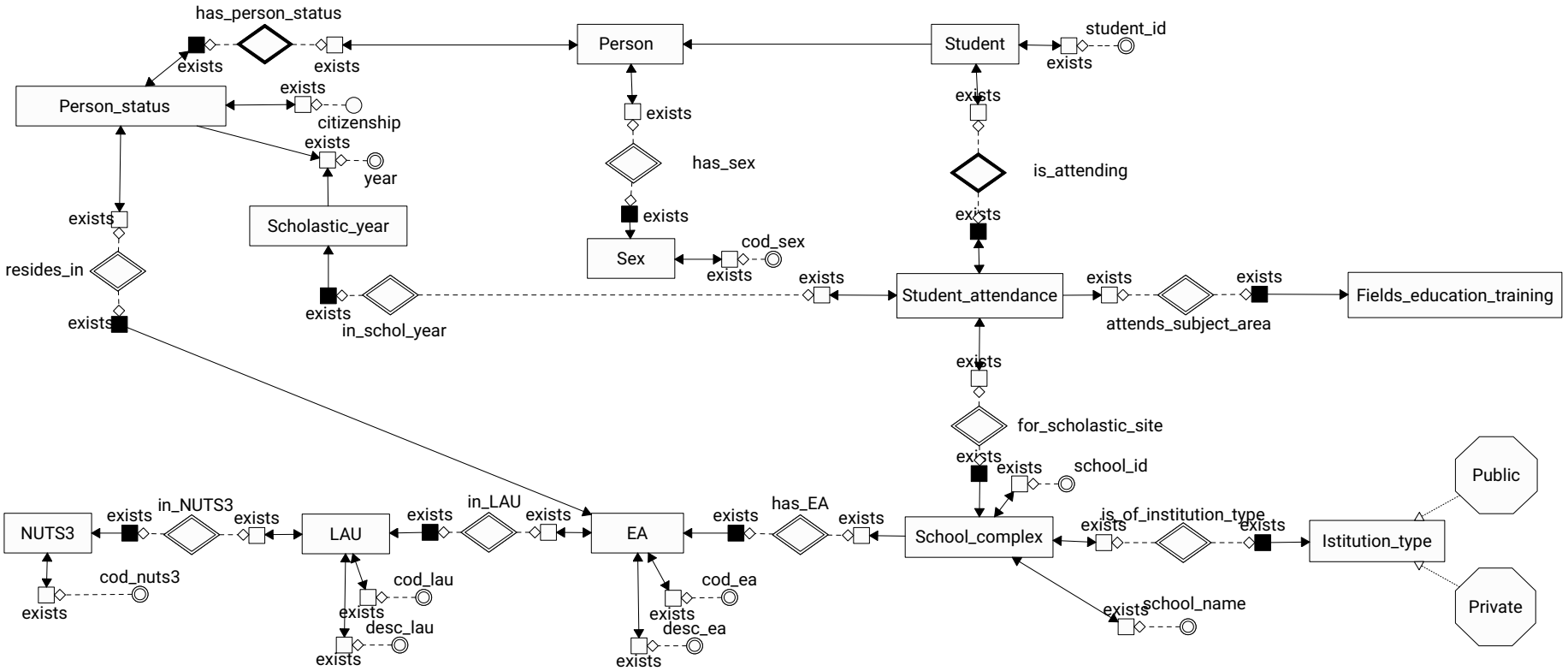
- Observation: the facts of the cube are populated starting from micro-data managed within the organization information system

  →within ISTAT, the facts instantiating the cube can be intensionally described by means of a query over the domain ontology, i.e., retrieving students attending a school in Italy since 2015

- Proposal

model macro-data by explicitly representing the relationship with micro-data they have been computed from

The INTERSTAT domain ontology



FUTURE AI RESEARCH

INTERSTAT Views definition

- **View** Attendance(id,year,sex,s_code,s_ea) **as**
  $(id,y,s,c,eac) :- student\_id(p, id)$, $has\_sex(p, s)$,
  $has\_person\_status(p, ss)$, $year(ss, y)$, $citizenship(ss,' Italian')$,
  $is\_attending(p, sa)$, $in\_schol\_year(sa, y)$, $for\_scholastic\_site(sa, sc)$,
  $school\_id(sc, c)$, $has\_EA(sc, ea)$ $cod\_ea(ea, eac)$, $y > 2015$

- **View** enumToLocal(eArea,locUnit) **as** $(e, l) : -in\_LAU\ (e, l)$

- **View** localToTerr(lUnit,terrUnit) **as** $(l,t) : -in\_NUTS3(l,t)$

INTERSTAT hierarchies and cubes definition

- **Hierarchy** HSpace **with edges**
  { (eArea,enumToLocal,locUnit), (locUnit,localToTerr,terrUnit) }

- **Base Data Cube** *BDC*1 **on view** Attendance **with dimensions**

  scholYear **from** year
  sex **from** sex
  location **from** s_ea **with hierarchy** HSpace

  with measures $count()$ **as** *qty*

- **Data Cube** *DDC*1 **on cube** $BDC1$ **Roll-up on dimension**

  sex

  location **at node** terrUnit **of hierarchy** HSpace

  with measures Sum**(qty) as qty**

Conclusions and future work

- We have formalized the approach proposed by introducing the notion of multidimensional ontology, including both views and cubes definitions
- We based our approach on the Metamodeling Semantics proposed in [Lepore et al, AIJ 2021]
- We paved the way to investigate and realize reasoning services to enable comparisons among cubes

* Maurizio Lenzerini, Lorenzo Lepore, Antonella Poggi: Metamodeling and metaquerying in OWL 2 QL. Artif. Intell. 292: 103432 (2021)