

# Bridging the gap between micro and macro data: Ontologies to the rescue

Domenico Lembo<sup>1</sup>, Maurizio Lenzerini<sup>1</sup>, Antonella Poggi<sup>1</sup>, Roberta Radini<sup>2</sup>,  
Michele Riccio<sup>2</sup> and Valerio Santarelli<sup>3</sup>

<sup>1</sup>Sapienza Università di Roma

<sup>2</sup>Istituto Nazionale di Statistica (ISTAT)

<sup>3</sup>OBDA Systems s.r.l.

## Abstract

We describe a new methodology for modeling aggregate data and explicitly connecting them to the individual-level data from which aggregates are generated. The approach makes use of OWL2 ontologies that formalize both the application domain and multidimensional constructs, such as data cubes, measures, dimensions, and hierarchies. This contribution stems from a collaboration among ISTAT, Sapienza University of Rome, and OBDA Systems, within the project INTERSTAT.

## Keywords

Multidimensional Data Modeling, Ontologies, Data Warehousing

## 1. Introduction

Aggregate data, also known as macro-data, concern with information produced in summarized form from individual level data, also known as micro-data. Typically gathered from operational databases, and possibly integrated from various data sources, aggregate information is usually managed through Business Intelligence and Data Warehousing solutions [1]. It is often included into reports or dashboards, and used to support decision-making processes within an organization. Moreover, aggregate data may be distributed by organizations (e.g., statistical or other institutional bodies) in the form of open data, freely exploitable by external stakeholders.

Data aggregation is usually carried out by referring to the so-called *multidimensional model* [2], where events of interest for the analysis are represented as logical cubes. These Cubes are characterized by *dimensions*, which correspond to the aspects of the business along which one wants to perform aggregation (e.g., time or space), possibly associated to *hierarchies* specifying different *levels* of aggregation (also known as dimensional attributes [2]), and by *measures*, which are properties of the event on which to make calculations (e.g., sums or averages) and that can be used as business performance indicators (e.g., income of a shop, number of enrollments in a school). Operations performed on data cubes (also called OLAP operations) include the

---


SEBD 2023: 31st Symposium on Advanced Database System, July 02–05, 2023, Galzignano Terme, Padua, Italy

✉ lembo@diag.uniroma1.it (D. Lembo); lenzerini@diag.uniroma1.it (M. Lenzerini); poggi@diag.uniroma1.it (A. Poggi); radini@istat.it (R. Radini); riccio@istat.it (M. Riccio); santarelli@obdasystems.com (V. Santarelli)

🆔 0000-0002-0628-242X (D. Lembo); 0000-XXXX-XXXX-XXXX (M. Lenzerini); 0000-0002-4030-3458 (A. Poggi)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

increment or decrement of the level of aggregation, called roll-up and drill-down, respectively, or the selection of a portion of events in the multidimensional space, called slice-and-dice.

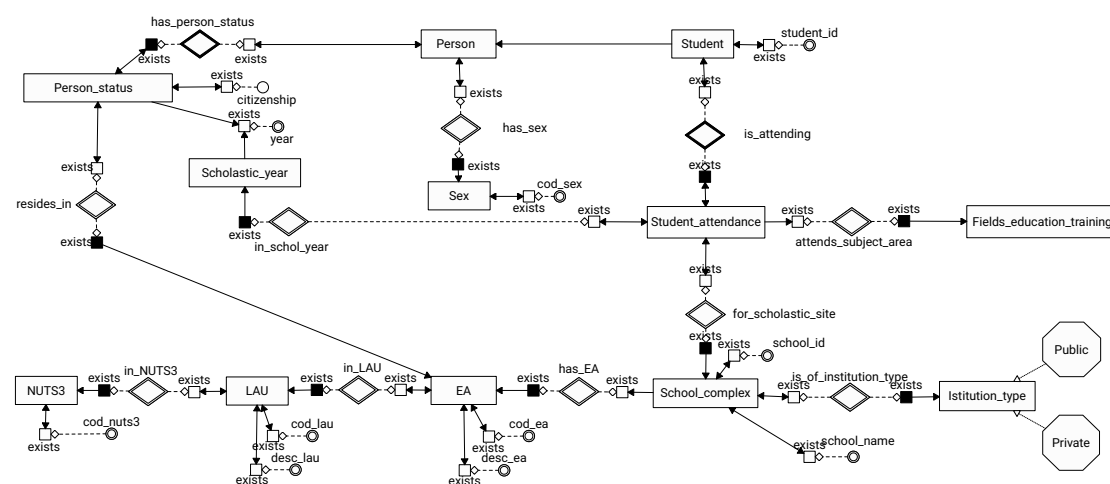
In this paper we propose a new methodology for modeling and manipulating aggregate data, which is based on the use of OWL2 ontologies that provide a rigorous formalization of both the application domain and the multidimensional model. The overall ontology that we devise makes it explicit the way in which macro-data are obtained from micro-data, by exploiting *views* over the domain ontology, which are first-class citizens in our model. Data cubes and hierarchies are indeed seen as constructed from the (SPARQL) queries associated to the views, which allow cubes dimensions, cubes measures and hierarchy levels, to be instantiated from the answers to such queries. This is a distinguishing feature of our approach, considered that other models for multidimensional data (e.g., [3, 4]) do not formalize this aspect, and methodologies for data warehouse design do not provide declarative means to specify the connection between micro- and macro-data, which is usually hidden in ETL procedures, and thus it is difficult to understand and reconstruct, e.g., for data provenance and/or lineage.

We remark that our ontology is equipped with a tailored higher-order semantics, in the spirit of [5]. This paves the way for developing advanced reasoning services, as, e.g., processing queries that mix together meta-level categories (as cubes or hierarchies) and domain elements. Such an aspect is particularly interesting considered also the possibility of linking the ontology to both micro-data repositories and aggregate data sets through mappings, thus extending the Ontology-based Data Management (OBDM) [6] paradigm to the presence of multidimensional data. Finally, we point out that our approach enables for the understanding and integration of aggregate data, possibly produced independently by different organizations.

## 2. The INTERSTAT project

Since 2015, ISTAT has developed the Integrated System of Statistical Registers (ISSR), which is a set of registers distinguished according to the following themes: socio-demographic, territorial, and institutional. Data of the different registries have been integrated and made interoperable through ontologies. The application of the OBDM methodology to the ISSR has made it possible to create a single conceptual access point to the above mentioned information assets, and more generally to obtain data governance of the entire micro-data system of the Institute [7, 8, 9]. Analogous needs do in fact emerge also for the management of macro-data interoperability. Macro-data represent the core business of statistical institutes and often are at the basis of a significant part of the information services of many public bodies, which often collect such data from various institutional or private producers. The INTERSTAT project studies how to achieve interoperability of statistical data from a cross-border and cross-domain point of view. In this respect, one of the findings of the project is that the relationship between micro- and macro-data should be clearly formalized. To this aim, leveraging the experience of ISSR, within the project we realized an overall ontology that makes such relationship explicit, thus allowing for the harmonization of aggregated and non-aggregated data.

The project carried out three pilots to test semantic interoperability methods on different topics and with respect to data producers with distinct information purposes. The final goal of each pilot is to create synthetic indicators to support public decision-makers.



**Figure 1:** The school domain ontology represented in the Graphol language [10]

### 3. Illustrating the approach

We illustrate our approach through an example that refers to the “School for You” (S4Y) pilot of INTERSTAT. This pilot concerns with the integration, from various sources, of data related to school attendance in Italy and France, for the construction of comparative indicators on the population of students by order of study. After presenting the domain ontology, which provides a conceptual representation of the domain of interest, we describe how, starting from the ontology, we have defined a set of views and then, based on such views, a set of hierarchies and cubes, to support analytical tasks regarding specific phenomena. The final result is an overall ontology providing a unified view over the domain of interest (i.e., the representation of *micro-data*) and the domain of views, hierarchies and cubes defined for its analysis (i.e., the conceptualization of the *macro-data*).

**The domain ontology.** The Ontology describes the domain of interest in terms of concepts (also known as classes), roles (also known as object properties), and attributes (also known as data properties). Specifically, it represents persons (concept `Person`) in terms of some of their features, such as sex, birth date, citizenship and residence. It then describes students (concept `Student`), which are persons who have a student id and attend schools in some scholastic years. In particular, the school attendance by students (concept `Student_attendance`) is characterized by the study subject area and the school complex (concept `School_complex`) where the student is registered. Each school complex can be a public or private institution, and is described in terms of its identifying code, name, and Enumeration Area where it is located (concept `EA`), which belongs to a local administrative unit (concept `LAU`). Finally, each local administrative unit (concept `NUTS3`) is part of a territorial unit.

We point out that, since the person citizenship and residence may vary during the years and we are interested in keeping track of the variations, we assume to represent the status of a person  $p$  at the beginning of each year, by associating to  $p$  an instance  $sp$  of the concept `Person_status` that is characterized by the attributes `year` and the `citizenship`, and by

id	year	sex	s_code	s_ea
10029	2020	male	sc24	201
10029	2021	male	sc25	204
10029	2022	male	sc25	204
15442	2020	male	sc24	201
15442	2021	male	sc25	204
12024	2017	female	sc12	32
12024	2018	female	sc25	204
22378	2018	female	sc25	204
52627	2018	female	sc25	204
34567	2017	female	sc15	47
01023	2017	female	sc15	47

**Figure 2:** Attendance view extension

the Enumeration Area where the person resides (cf. `role resides_in`).

**Definition of the views.** As an example, we here discuss analyses over aggregate data about school attendance of Italian students, both male and female, since 2015. All relevant information is scattered through the ontology. In fact, our analyses involve several domain concepts, playing the role of *multiple statistical units*, namely the attendance, the students and the school addresses, which are related to each other through a specific set of conditions. This is one of the circumstances in which, in our approach, we resort to views, which allow to formally capture through a query over the domain ontology all relevant data of multiple statistical units. Indeed, once appropriate views are defined, we can use them to specify aggregates for analytical tasks. Note that, as we will see in the next subsection, views over the ontology are also used to define hierarchies allowing to navigate data through different aggregation levels.

We thus define the view `Attendance` as follows:

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

The target variables of the query (e.g.  $y$  or  $s$ ) defining the view are in one-to-one correspondence with the *view attributes* (e.g., `year,sex`). These attributes refer to all and only the data of interest for a specific *set* of investigations. For instance, the `Attendance` view does not include the study subject area nor the school type of institution, which are not needed the analyses at hand.

The extension of the view, executed over the ontology under certain answers semantics [11], is shown in Fig. 2.

**Hierarchies.** As already mentioned, analytical tasks often require to aggregate data at different levels of granularity. This is achieved by means of built-in or customized *hierarchies*. In our approach, hierarchies are defined in terms of the domain ontology, by exploiting views. In more detail, we define a hierarchy  $h$  by specifying its *intension* as the set of pairs of *nodes* (i.e., the

eArea	locUnit
201	l1
204	l1
32	l2
47	l3

(a) enumToLocal

locUnit	terrUnit
l1	t1
l2	t2
l3	t2

(b) localToTerr

**Figure 3:** Extension of views associated to edges of HSpace**Figure 4:** Graphical representation of hierarchies

hierarchy levels) constituting the edges of a Directed Acyclic Graph (DAG), where each edge is associated to a binary view. Intuitively, the *extension* of the hierarchy is another DAG whose edges are the pairs of values in the view extension.

Turning the attention to our example, in order to be able to navigate data aggregated on the basis of different levels of territorial partitioning, we first define two views `enumToLocal` and `localToTerr` consisting of all pairs  $(e, l)$  such that  $e$  is an enumeration area belonging to the local unit  $l$  and of all pairs  $(l, t)$  such that  $l$  is a local unit belonging to the territorial unit  $t$ , respectively:

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

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

The extensions of `enumToLocal` and `localToTerr` are shown in Fig. 3.

By exploiting the views above, we define the hierarchy named `HSpace` as follows:

#### Hierarchy HSpace with edges

$\{(eArea, enumToLocal, locUnit),$   
 $(locUnit, localToTerr, terrUnit)\}$

where `eArea`, `locUnit`, `terrUnit` are the three nodes of the hierarchy `HSpace`, whose intension is the DAG graphically depicted in Fig. 4a, and whose extension is the DAG depicted in Fig. 4b.

**Base Data Cube.** Suppose that the primary events of interest for our analysis refer to the number of Italian students who attended a school in Italy since 2015, per scholastic year, sex, and school address. We thus define a *base* data cube as follows:

scholYear	sex	location	qty
2020	male	201	2
2021	male	204	2
2022	male	204	1
2017	female	32	1
2018	female	204	3
2017	female	47	2

(a)  $BDC_1$  extension

sex	location	qty
male	t1	5
female	t2	3
female	t1	3

(b)  $DDC_1$  extension**Figure 5:** Examples of data cubes**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*

The above definition specifies that (i)  $BDC_1$  is defined on the view *Attendance*, (ii) that it has dimensions *scholYear* from (the view attribute) *year*, *sex* from *sex*, and *location*, from *s\_ea* with hierarchy *HSpace*, and, finally, (iii) that it counts the number of tuples in the view having the same values for *year*, *sex*, and *s\_ea*; this measure is named *qty* (the operator used to compute it is *count()*).

Given the extension of *Attendance* in Fig. 2,  $BDC_1$  is instantiated by the *observed events* shown in Fig. 5a.

Note that  $BDC_1$  projects some components of the view *Attendance* out, in particular those that do not play any role in the data cube. As said, views are typically designed for a *set* of analyses, and indeed *Attendance* is at the basis of the definition of other cubes.

**Derived Data Cubes.** Once we have defined a base data cube, we may want to represent derived data cubes obtained from the base one (or from other derived data cubes) by applying OLAP operators such as Roll-up, Drill-down and Slice and Dice. For the lack of space, we next illustrate only the case of Roll-up.

A data cube is obtained by another through a roll-up by specifying the wanted aggregation level along one or more (hierarchies associated to) dimensions. For example, suppose that we want to apply the roll-up operator to the (base) data cube  $BDC_1$ , to get the data cube  $DDC_1$  reporting the number of Italian students who attended a school in Italy since 2015, per sex and per territorial unit. To this aim we use the following specification:

**Data Cube  $DDC_1$  on cube  $BDC_1$** **Roll-up on dimension***sex**location* **at node** terrUnit **of hierarchy** HSpace**with measures** *Sum(qty)* **as** *qty*

The above definition states that  $DDC_1$  is the result of applying the Roll-up operator to  $BDC_1$ , towards the *terrUnit* node of the hierarchy *HSpace*, and by eliminating the *scholYear* dimension,

which does not appear among the dimensions of  $DDC_1$  (note that here we are “rolling-up” the entire degenerate dimension *scholYear* [1]).

Given the extension of HSpace in Fig. 4b and of  $BDC_1$  in Fig. 5a, the extension of  $DDC_1$  is that shown in Fig. 5b.

## 4. Conclusion

Our work is currently focused on the development of services to support both design- and run-time activities related to the production, distribution and integration of aggregate data. Such services are defined according to a formal semantics that extends the Metamodeling Semantics proposed in [5]. This semantics allows to reason over the various representation layers of the overall ontology we realized, i.e., the meta-level formalizing the multidimensional model, the actual data cubes designed for the analysis of the business trends, the domain ontology and the views bridging it to the cubes. A fundamental service in this scenario is query answering. Such service is indeed at the basis of several more complex functionalities, such as integration of aggregate data sets, possibly acquired from external sources and suitably linked to the ontology through mappings as in OBDM, and production and publishing of linked open data. Interestingly, queries in our framework may smoothly combine together elements belonging to the various levels of the ontology. This allows, for instance, to pose a query as the following

$$q(x, y) \quad :- \quad \text{Cube}(x), \text{Cube}(y), \text{based\_on}(x, v_1), \\ \text{based\_on}(y, v_2), \text{disjoint}(v_1, v_2)$$

In words, the above query is asking for all pairs of cubes that are based on disjoint views (i.e., views without common answers), and can thus be considered incomparable. Notice that this querying ability goes beyond those of current systems that manage aggregate information. Our efforts are thus concentrated in the implementation of software components, possibly integrated in the OBDM tool Mastro [12], that realize the above idea.

## References

- [1] R. Kimball, M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3 ed., Wiley, 2013.
- [2] M. Golfarelli, S. Rizzi, *Data Warehouse Design: Modern Principles and Methodologies*, McGraw-Hill, 2009.
- [3] R. Cyganiak, D. Reynolds, *The RDF Data Cube Vocabulary*, W3C Recommendation, W3C, 2014. Available at <https://www.w3.org/TR/vocab-data-cube/>.
- [4] The official site for the SDMX community, <https://sdmx.org/>, 2023.
- [5] M. Lenzerini, L. Lepore, A. Poggi, Metamodeling and metaquerying in OWL 2 QL, *AIJ* 292 (2021) 103432.
- [6] M. Lenzerini, Managing data through the lens of an ontology, *AI Magazine* 39 (2018) 65–74.
- [7] R. Aracri, A. M. Bianco, M. Scannapieco, L. Lepore, R. Radini, V. Santarelli, L’uso delle ontologie per la governance dei dati del sir, in: *Ital-IA*, 2019.

- [8] R. M. Aracri, A. Bianco, R. Radini, M. Scannapieco, L. Tosco, Using ontologies for official statistics: The istat experience, in: *Practi-o-Web*, 2017.
- [9] R. Radini, M. Scannapieco, G. Garofalo, The italian integrated system of statistical registers: On the design of an ontology-based data integration architecture, in: *NTTS*, 2017.
- [10] D. Lembo, V. Santarelli, D. F. Savo, G. D. Giacomo, Graphol: A graphical language for ontology modeling equivalent to OWL 2, *Future Internet* 14 (2022).
- [11] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, R. Rosati, Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family, *J. of Automated Reasoning* 39 (2007) 385–429.
- [12] Mastro - The OBDM Engine, <https://obdm.obdasystems.com/mastro/>, 2023.