Le ontologie per l'accesso ai dati e l'interoperabilità semantica

Antonella Poggi – Sapienza University of Rome
Valerio Santarelli – OBDA Systems Srl

Department of Computer, Control, and Management Engineering Antonio Ruberti

Sapienza
Università di Roma

obda
SYSTEMS
Almawave Group

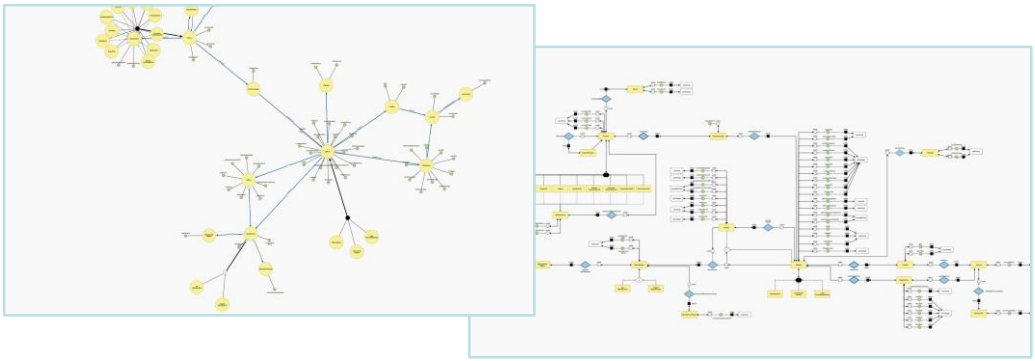
Outline

1. Le ontologie, l’OBDM e i suoi vantaggi
2. L’OBDM in Istat con Monolith
3. Interstat: il progetto e i suoi risultati

Le ontologie, l’OBDM e i suoi vantaggi

Ontology-based Data Management

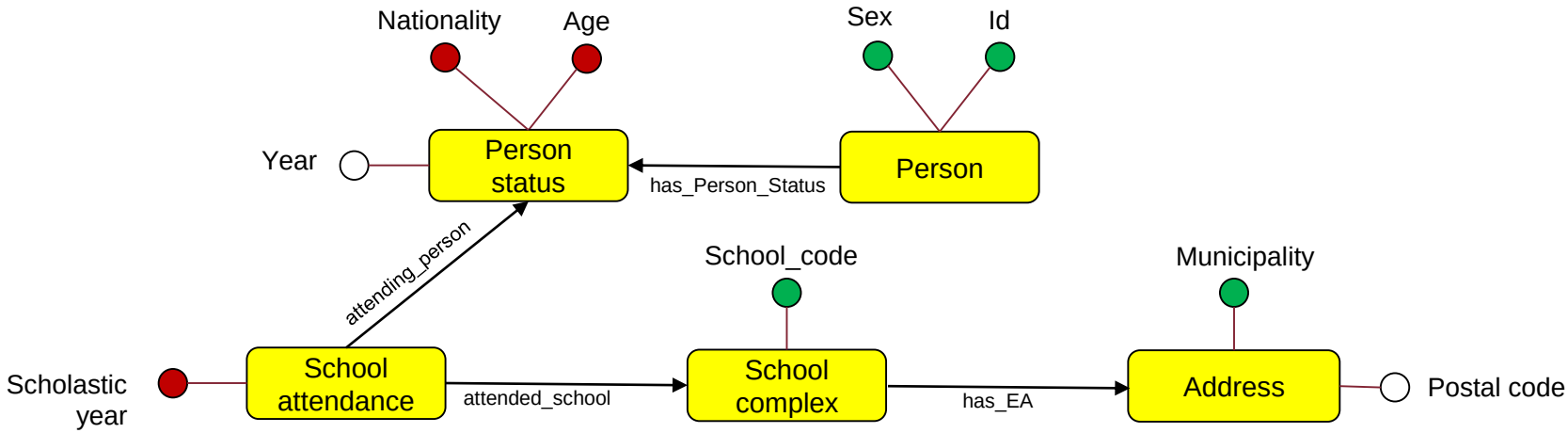
Tecnologia di **Data virtualization** che permette l'**accesso ai dati** attraverso un'**Ontologia** o un **Knowledge Graph (KG)**



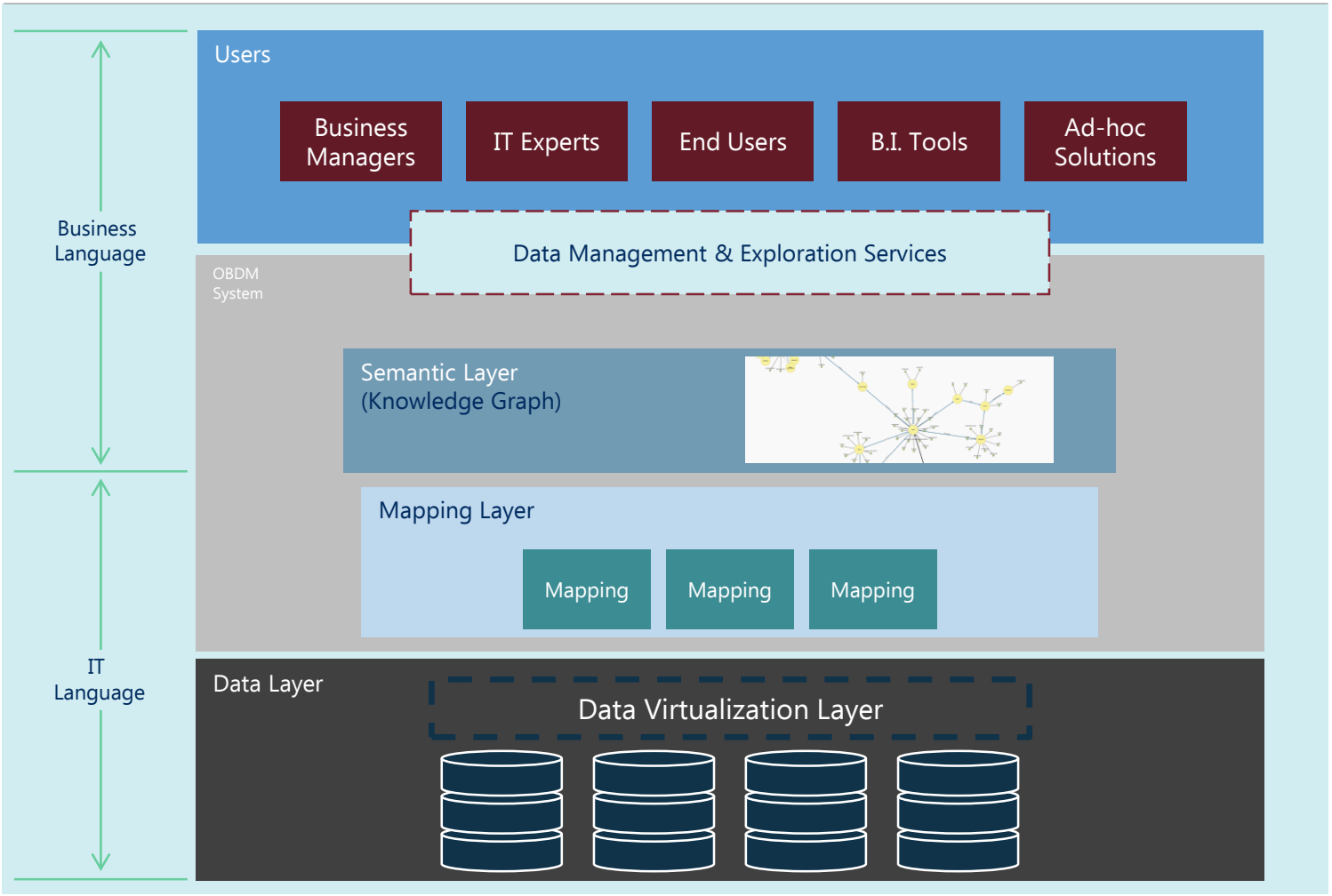
- Ontologie e KG sono modelli di dati semantici, che rappresentano i dati e il loro significato sotto forma di grafo
- Il grafo offre uno strato semantico tra i dati e l'utente che:
  - organizza i dati evidenziando relazioni e collegamenti
  - semplifica la formulazione dei requisiti informativi
  - facilita la comprensione delle risposte.

Interrogare un’ontologia

- Esempio: ”Informazioni (id,età,sesso,codice scuola,area scuola) sugli studenti di nazionalità italiana che hanno frequentato nell’a.s. 2014-15 una scuola a Roma"
- Grafo della query:



Architettura di un sistema di OBDM



Servizi offerti e Vantaggi

Integrazione

Esplorazione

Verifiche di Qualità

Self-service B.I

Astrazione

Chiarezza

Agilità

Incrementalità

Semplicità

Efficienza

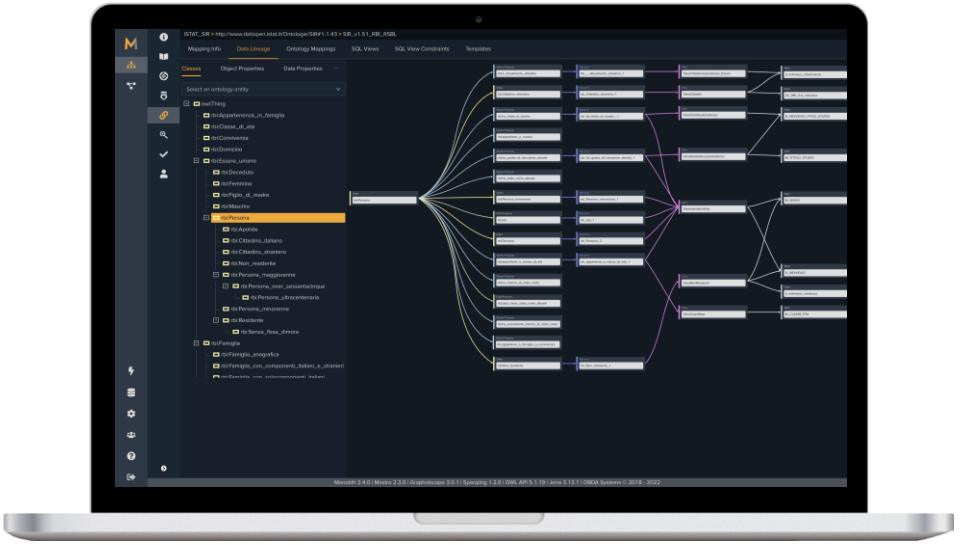
L’OBDM in Istat con Monolith

MONOLITH

Monolith è una piattaforma che abilita la gestione semantica di dati aziendali tramite ontologie e knowledge graphs, permettendo di erogare i servizi dell'OBDM attraverso Mastro, il suo motore di ragionamento su ontologie.

Costruzione della specifica

- Ontology Designer
- SQL-based Mapping Designer



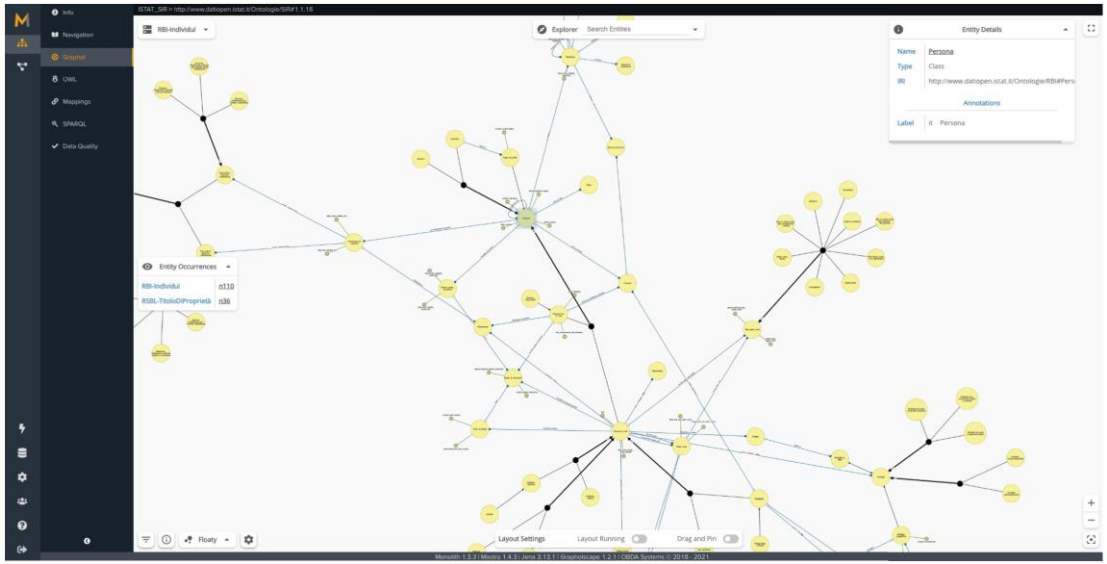
Servizi sui dati

- Esplorazione e navigazione del modello e dei dati
- No-code query builder
- Data Lineage
- Costruzione di processi di Data Quality
- Trasformazione di dati relazionali in graph data
- Integrazione con dataset L.O.D.

SIR – Sistema Integrato dei Registri Statistici

Il Progetto: Gestione dei dati del SIR attraverso OBDM

- Progetto nato da una collaborazione tra l'ISTAT, la Sapienza Università di Roma, ed OBDA Systems
- Il SIR è il Sistema di Registri Statistici: centralizza ed integra i dati derivati dalle fonti amministrative e dalle indagini statistiche condotte dall'Istituto
- Sistema integrato dei dati che fanno riferimento ai tre *tematismi* principali delle statistiche: popolazione, unità produttive, e territorio



SIR – Sistema Integrato dei Registri Statistici

L’Ontologia del SIR

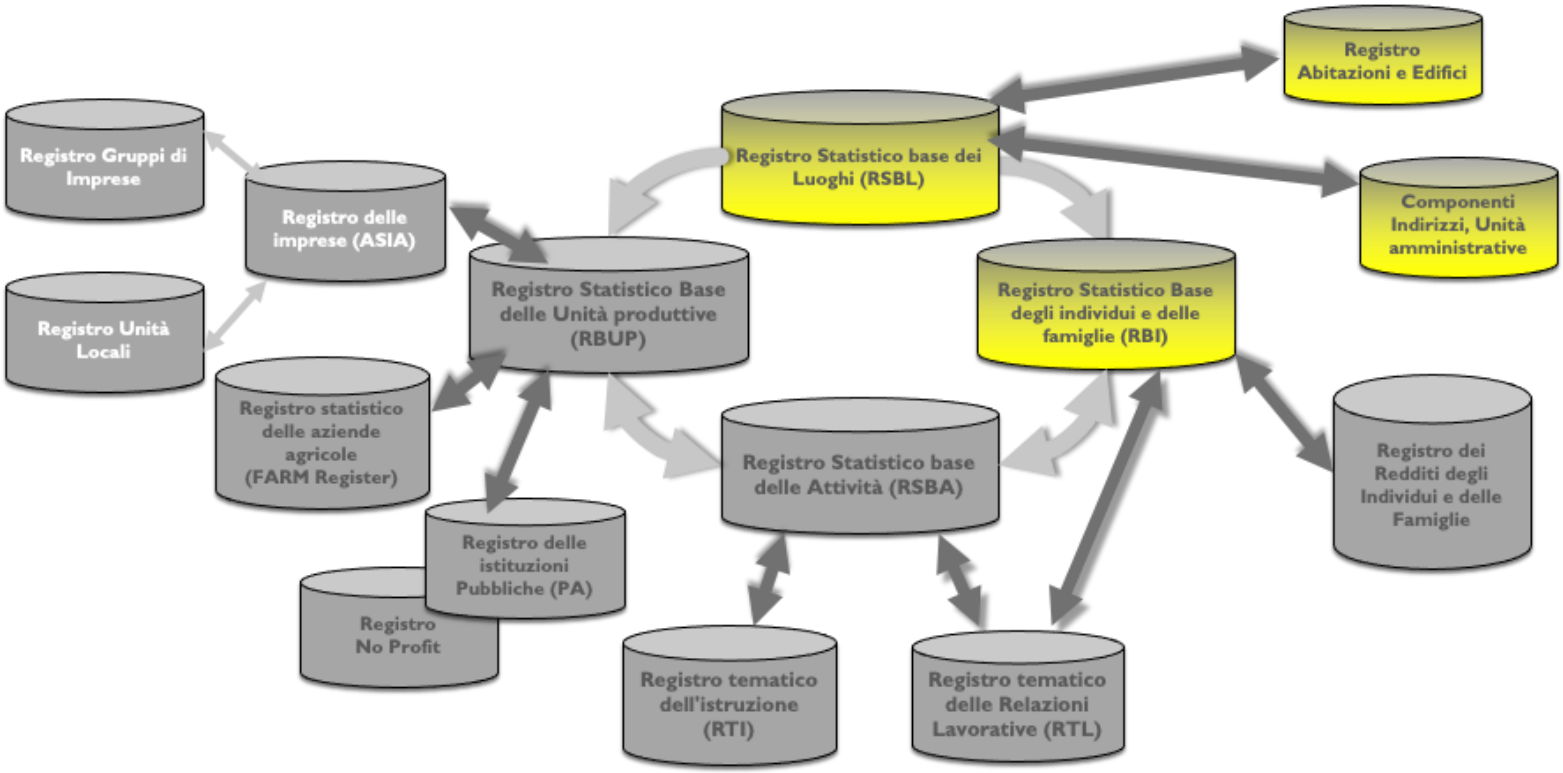
- Integrazione e Standardizzazione
  - Unificare il patrimonio informativo dei registri all’interno di una visione concettuale di alto livello
  - Standardizzare vocabolari e terminologia inter-registro

Monolith e l’OBDM

- Self-Service Data Access
  - Accesso in modo diretto ai dati e metadati del SIR attraverso query *no-code* o NL sull’ontologia
- Data Quality
  - Verifiche di qualità del dato con obiettivo di bonificare le sorgenti dei dati
- R&D
  - Sviluppo di nuove soluzioni per la gestione semantica della conoscenza attraverso attività congiunta di R&D

SIR – Sistema Integrato dei Registri Statistici

So far...



NDC - National Data Catalogue

Il titolare del progetto è il Dipartimento per la trasformazione digitale, l'Istituto Nazionale di Statistica né è l'attuatore

Obiettivi del progetto:

- Modellazione semantica e armonizzazione informativa delle procedure e dei servizi della pubblica amministrazione per la progettazione di servizi digitali interoperabili a livello nazionale e internazionale nell'ambito degli Stati membri europei
- Produzione di modelli semantici e standard comuni per scambio dei dati e comprensione delle informazioni

Risorse

- Ontologie, Vocabolari Controllati e Schemi di Dati

Attività in corso

- Aggiornamento delle ontologie core pubblicate sul portale *schema.gov.it,* per adeguarle alle esigenze di rappresentazione concettuale dei domini applicativi delle PPAA
- Creazione di nuove ontologie dei domini applicativi delle PPAA significativi per il progetto NDC
- Definizione di vocabolari controllati a supporto della standardizzazione semantica

National Data Catalog: Risultati Raggiunti e Prospettive

Il catalogo contiene decine di Ontologie e Vocabolari controllati che modellano gli asset semantici di diverse organizzazioni e cresce grazie al continuo contributo e partecipazione delle PPAA al progetto.

Da Settembre 2022 ad oggi stanno contribuendo:
Ministero degli Interni, Ministero dell'Università e della Ricerca,
Istituto Nazionale Assicurazione Infortuni sul Lavoro e Istituto Nazionale della Previdenza Sociale

La crescita dei contenuti semantici:

- Piu 1500 Nuove entità modellate
- 47 Nuove ontologie e vocabolari controllati realizzati

Attività Future e orizzonte sui dati

- Analisi e sviluppo nuove ontologie, vocabolari ed di schemi di e-service per procedure, anche SDG, ad esempio: Dichiarazione dei redditi di impresa e di pensioni fisiche, Immatricolazioni, Cambi indirizzo, …
- Interrelazioni con **PDND**[1] per la realizzazione dell'interoperabilità dei sistemi informativi e delle basi di dati delle PPAA attraverso le API (e-service)
- Riutilizzo delle risorse semantiche di NDC per la pubblicazione di dataset LOD

Possibili applicazioni del OBDM: NDC un patrimonio semantico da sfruttare

- la mappatura attraverso l'OBDM delle ontologie di NDC sulle sorgenti dati delle PPAA afferenti (anche quelle che non partecipano al progetto) abiliterebbe l'accesso semantico al dato, secondo modelli standardizzati e condivisi
- la realizzazione della consistenza semantica dei dati delle diverse Basi Dati delle diverse PPAA
- Le tecniche del OBDM possono contribuire alla realizzazione dei Linked Open Data pubblicati come High Value Dataset

1 - https://innovazione.gov.it/argomenti/pdnd/

Interstat: il progetto e i suoi risultati

Il progetto INTERSTAT

Gli enti statistici nazionali collezionano, ognuno, dati **aggregati**, anche chiamati **macro-dati**, che provengono a loro volta da enti pubblici tra loro autonomi e che consentono analisi multidimensionali a scopi diversi

→ come derivare indicatori sintetici che siano di supporto ai decision-makers? E.g., Ministero della Pubblica Istruzione

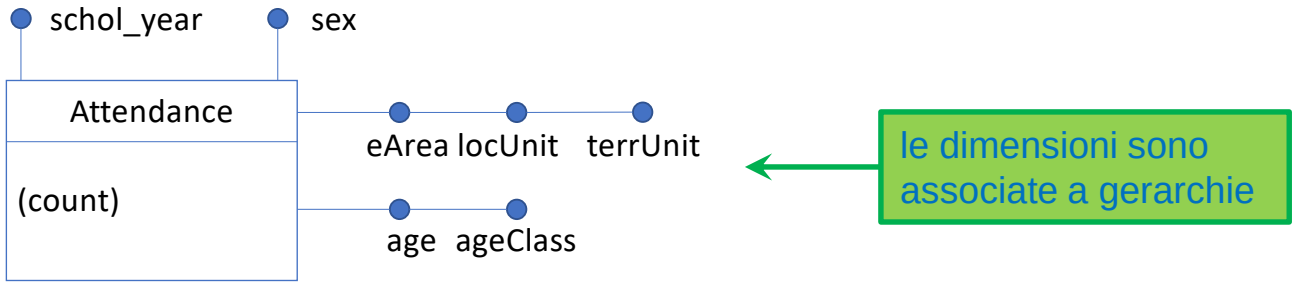
→ necessità di integrare macro-dati per consentire un'analisi multidimensionale unificata

Progetto pilota "School for You" (S4Y)

- Obiettivo: definire degli indicatori comparativi sulla popolazione degli studenti partendo da dati aggregati sulla frequenza scolastica in Italia e in Francia

S4Y: esempio di insieme di dati aggregati

- Numero di studenti che hanno frequentato una scuola in Italia dal 2015, classificati per
  - anno scolastico,
  - sesso,
  - localizzazione geografica delle scuole → i.e., classificate secondo un meccanismo standard usato per il censimento:
    - ad ogni scuola è associata una area di enumerazione
    - le aree di enumerazione sono associate alle unità amministrative locali
    - le unità amministrative locali sono associate alle unità territoriali al terzo livello della classificazione NUTS, che corrispodono alle province in Italia e ai «départements» in Francia
- Secondo il Dimensional Fact Model (DFM), l'insieme di dati aggregati sopra descritto si rappresenta come il seguente cubo



→ Ogni fatto che istanzia il cubo (chiamato anche evento) rappresenta la frequenza scolastica di un insieme di studenti charatterizzati da un anno scolastico, il sesso, e il luogo location, che può corrispondere ad un'area di enumerazione, un'unità locale o territoriale

S4Y: esempio di problema

- La definizione dello schema del cubo non esplicita che si riferisce a tutti e soli gli studenti che hanno frequentato una scuola in Italia dal 2015
  - metadati di questo tipo sono spesso descritti in maniera informale nella documentazione
  - per avere garanzie di completezza e correttezza sui metadati riguardanti i cubi, bisognerebbe analizzare il codice dei processi ETL (Extract – Transform – Load) che sono stati usati per ottenerli

→ Come facciamo a sapere se ha senso confrontare il cubo sulla frequenza scolastica in Italia con un cubo con dati «simili» riguardanti la frequenza scolastica in Francia?

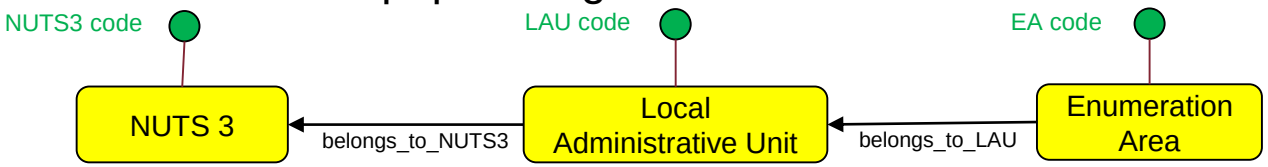
  - bisognerebbe poter inferire, dalla modellazione dei dati aggregati, che i due cubi si riferiscono allo stesso periodo e a popolazioni che sono selezionate secondo gli stessi criteri (i.e., sono tutti gli studenti o solo quelli di nazionalità italiana?)

Risultati del progetto

- Abbiamo definito un linguaggio e una metodologia per modellare, all'interno dell'ontologia stessa, i dati aggregati, esplicitando **la relazione che li lega ai micro-dati da cui sono stati calcolati**

Metodologia

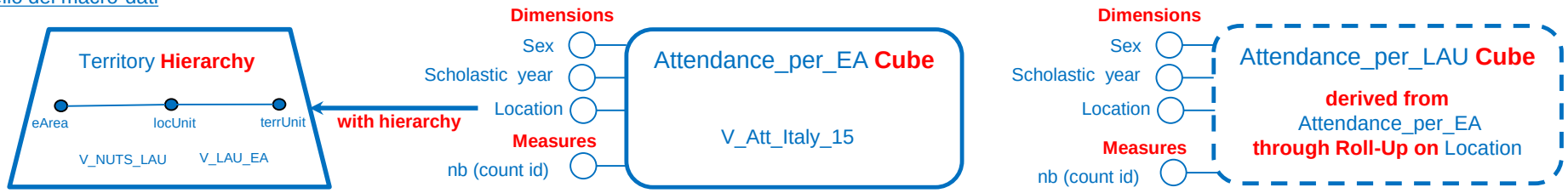
- Definiamo delle viste sull'ontologia, ovvero delle query che specificano, a livello intensionale, insiemi di dati che possono essere utili per l'analisi
  - come dati aggregati
  - come dati che popolano gerarchie



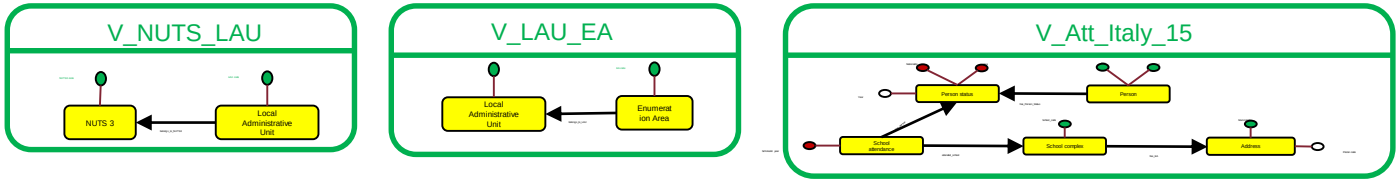
- Definiamo i cubi base a partire dalle viste
- Definiamo i cubi derivati a partire dai cubi base

Applichiamo la metodologia…

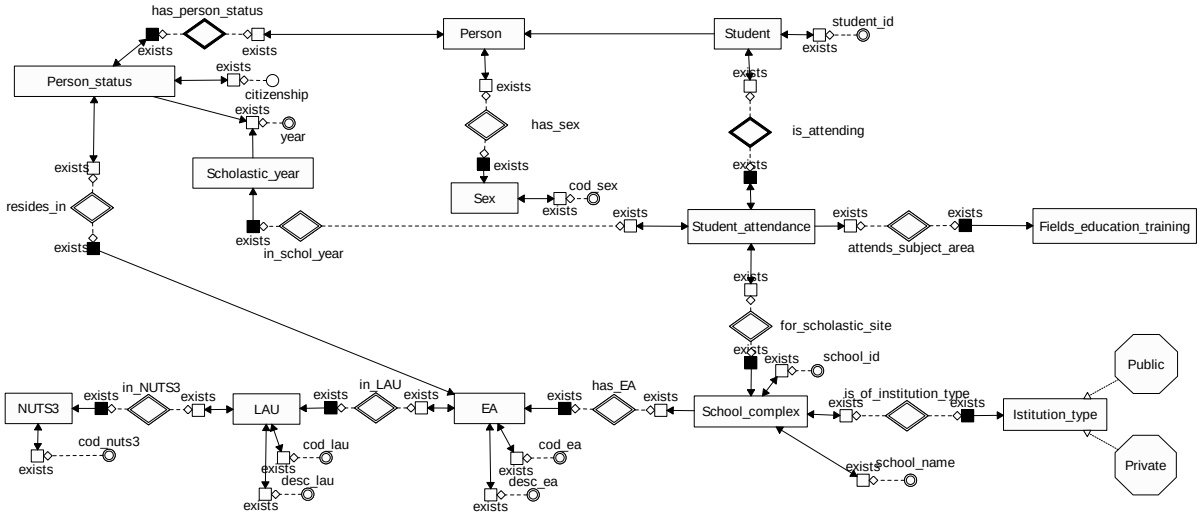
Livello dei macro-dati



Livello delle viste



Ontologia di dominio (dati micro)



Conclusioni e prospettive future

- Abbiamo definito le basi per consentire il confronto tra dati aggregati
- Abbiamo iniziato a studiare i servizi di ragionamento sull'ontologia estesa con i macro-dati
  - e.g., per il confronto tra insiemi di dati aggregati
- Abbiamo intenzione di collaborare con OBDA Systems Srl per far migrare i risultati del progetto in Monolith

Grazie!

Domande?

antonella.poggi@uniroma1.it
santarelli@obdasystems.com