An Integration-Oriented Ontology to Govern Evolution in Big Data Ecosystems

Sergi Nadal, Alberto Abelló, Oscar Romero
Universitat Politècnica de Catalunya, BarcelonaTech

Stijn Vansummeren
Université Libre de Bruxelles

Big Data ecosystems - What about Variety and Variability?

- Situational data (e.g., social networks) are mostly supplied by 3rd parties (e.g., Twitter) via REST APIs in semi-structured form
- Data analysts need to carefully study the documentation and adapt their tools to the particular schema provided.
- Endpoints are constantly evolving, hence analysts need to continuously adapt their tools to such changes

Ontology-Based Data Access (OBDA)

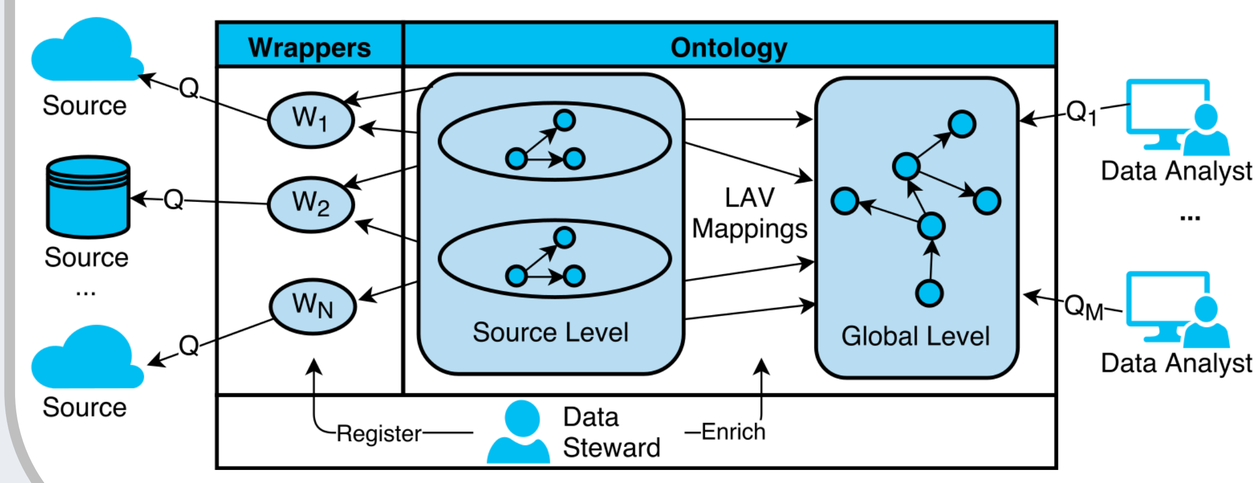
- Ontologies, a shared conceptualization of a domain
- Formalized by means of Description Logics (DLs)
- An ontology (*T*) is constituted by:
  - TBox: general properties of concepts and roles
  - ABox: instances of concepts and roles
- An OBDA system to query the ontology, and translate such queries to the sources (*S*) via mappings (*M*)
- TBox in *DL-Lite* and ABox in the sources, in original format

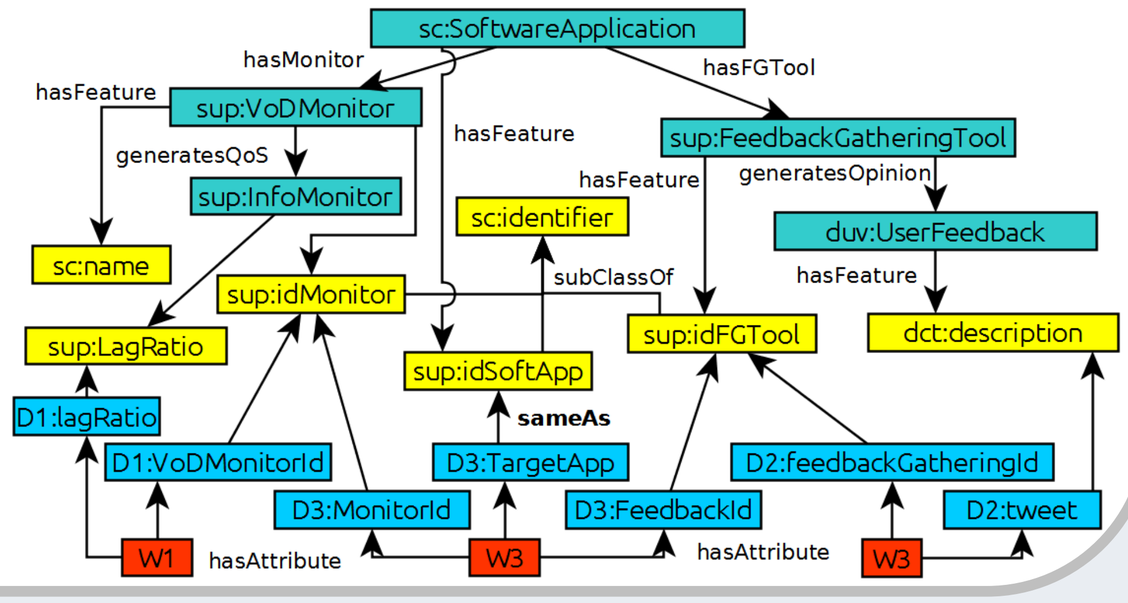
Why traditional OBDA is not enough?

- What if *S* changes? How to avoid queries on *T* crashing?
- Schema mappings follow the *global-as-view* approach
  - Elements of *T* are characterized as queries (views) over *S*
  - Simple query answering (unfolding), but changes in the sources might invalidate mappings
- We want *local-as-view* schema mappings
  - Elements of *S* are characterized as queries (view) over *T*
  - Loosely-coupling between the ontology and the sources, but query answering requires reasoning

The Big Data Integration Ontology

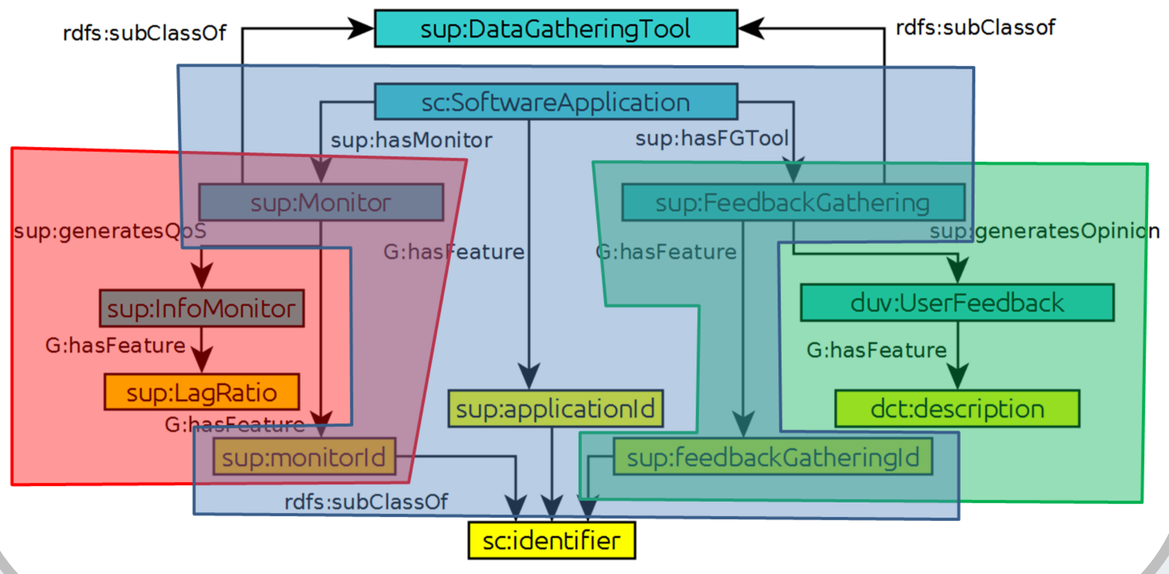
- We provide a new approach to OBDA with LAV mapping assertions to accommodate highly dynamic Big Data scenarios
- RDF vocabularies to annotate the elements of the ontology and drive the processes of evolution and query answering
- The Global level (*G*) depict concepts (green) and features (yellow), the Source lev. (*S*) wrappers (red) and schema (blue)
- The Data Steward responsible of registering wrappers of new or evolved sources, and creating LAV mappings





LAV Mappings

- Mappings are encoded as part of the ontology
- Each wrapper has associated a named graph over *G*, denoting what it is providing information about
  - A named graph is a subset of an RDF graph



Query Answering

- Given a SPARQL conjunctive query (CQ) over *G*, return an equivalent RA (union of CQs) over the wrappers

```
SELECT  ?w ?t
WHERE
  { ?t  rdf:type          sup:LagRatio .
    ?x  G:hasFeature      ?t ;
        rdf:type          sup:InfoMonitor .
    ?y  sup:generatesQoS  ?x ;
        rdf:type          sup:Monitor .
    ?z  sup:hasMonitor    ?y ;
        rdf:type          sc:SoftwareApplication ;
        G:hasFeature      ?w .
    ?w  rdf:type          sup:idApplication
  }
```

- The user can request only features, concepts do not exist in the sources but can be used to navigate
  - A query must be *well-formed*
- Concepts can only be joined via identifiers
- There exists many possible ways to combine the wrappers

Seventh European Business Intelligence & Big Data Summer School (eBISS 2017) - Brussels, Belgium