In Proc. of 2001 Int. Semantic Web Working Symposium **(SWWS 2001)**

A Framework for Ontology Integration

Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini
Dipartimento di Informatica e Sistemistica
Università di Roma "La Sapienza"
Via Salaria 113, 00198 Roma, Italy
{calvanese,degiacomo,lenzerini}@dis.uniroma1.it

Abstract. One of the basic problems in the development of techniques for the semantic web is the integration of ontologies. Indeed, the web is constituted by a variety of information sources, each expressed over a certain ontology, and in order to extract information from such sources, their semantic integration and reconciliation in terms of a global ontology is required. In this paper, we address the fundamental problem of how to specify the mapping between the global ontology and the local ontologies. We argue that for capturing such mapping in an appropriate way, the notion of *query* is a crucial one, since it is very likely that a concept in one ontology corresponds to a *view* (i.e., a query) over the other ontologies. As a result query processing in ontology integration systems is strongly related to view-based query answering in data integration.

1 Introduction

One of the basic problems in the development of techniques for the semantic web is the integration of ontologies. Indeed, the web is constituted by a variety of information sources, and in order to extract information from such sources, their semantic integration and reconciliation is required. In this paper we deal with a situation where we have various local ontologies, developed independently from each other, and we are required to build an integrated, global ontology as a mean for extracting information from the local ones. Thus, the main purpose of the global ontology is to provide a unified view through which we can query the various local ontologies.

Most of the work carried out on ontologies for the semantic web is on which language or which method to use to build the global ontology on the basis of the local ones [13, 2]. For example, the Ontology Inference Layer (OIL) [13, 2] proposes to use a restricted form of the expressive and decidable DL studied in [4] to express ontologies for the semantic web.

In this paper, we address what we believe is a crucial problem for the semantic web: how do we specify the mapping between the global ontology and the local ontologies. This aspect is the central one if we want to use the global ontology for answering queries in the context of the semantic web. Indeed, we are not simply using the local ontologies as an intermediate step towards the global one. Instead, we are using the global ontology for accessing information in the local ones. It is our opinion that, although the problem of specifying the mapping between the global and the local ontologies is at the heart of integration in the web, it is not deeply investigated yet.

We argue that even the most expressive ontology specification languages are not sufficient for information integration in the semantic web. In a real world setting, different ontologies are build by different organizations for different purposes. Hence one should expect the same information to be represented in different forms and with different levels of abstraction in the various ontologies. When mapping concepts in the various ontologies to each other, it is very likely that a concept in one ontology corresponds to a *view* (i.e., a *query*) over the other ontologies. Observe that here the notion of "query" is a crucial one. Indeed, to express mappings among concepts in different ontologies, suitable query languages should be added to the ontology specification language, and considered in the various reasoning tasks, in the spirit of [4, 5]. As a result query processing in this setting is strongly related to view-based query answering in data integration systems [20, 17]. What distinguishes ontology integration from data integration as studied in databases, is that, while in data integration one assumes that each source is basically a databases, i.e., a logical theory with a single model, such an assumption is not made in ontology integration, where a local ontology is an arbitrary logical theory, and hence can have multiple models.

Our main contribution in this paper is to present a general framework for an ontology of integration where the mapping between ontologies is expressed through suitable mechanisms based on queries, and to illustrate the framework proposed with two significant case studies.

The paper is organized as follows. In the next section we set up a formal framework for ontology integration. In Sections 3 and 4, we illustrate the so called global-centric approach and local-centric approach to integration, and we discuss for each of the two approaches a specific case study showing the subtleties involved. In Section 5 we briefly present an approach to integration that goes beyond the distinction between global-centric and local-centric. Finally, Section 6 concludes the paper.

2 Ontology integration framework

In this section we set up a formal framework for *ontology integration systems* (OISs). We argue that this framework provides the basis of an *ontology of integration*. For the sake of simplicity, we will refer to a simplified framework, where the components of an OIS are the global ontology, the local ontologies, and the mapping between the two. We call such systems "one-layered". More complex situations can be modeled by extending the framework in order to represent, for example, mappings between local ontologies (in the spirit of [12, 6]), or global ontologies that act as local ones with respect to another layer.

In what follows, one of the main aspects is the definition of the semantics of both the OIS, and of queries posed to the global ontology. For keeping things simple, we will use in the following a unique semantic domain Δ, constituted by a fixed, infinite set of symbols.

Formally, an OIS $\mathcal{O}$ is a triple $\langle \mathcal{G}, \mathcal{S}, \mathcal{M}_{\mathcal{G},\mathcal{S}} \rangle$, where $\mathcal{G}$ is the global ontology, $\mathcal{S}$ is the set of local ontologies, and $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ is the mapping between $\mathcal{G}$ and the local ontologies in $\mathcal{S}$.

Global ontology. We denote with $\mathcal{A}_\mathcal{G}$ the alphabet of terms of the global ontology, and we assume that the global ontology $\mathcal{G}$ of an OIS is expressed as a theory (named simply $\mathcal{G}$) in some logic $\mathcal{L}_G$.

Local ontologies. We assume to have a set $\mathcal{S}$ of n local ontologies $\mathcal{S}_1, \ldots, \mathcal{S}_n$. We denote with $\mathcal{A}_{\mathcal{S}_i}$ the alphabet of terms of the local ontology $\mathcal{S}_i$. We also denote with $\mathcal{A}_\mathcal{S}$ the union of all the $\mathcal{A}_{\mathcal{S}_i}$'s. We assume that the various $\mathcal{A}_{\mathcal{S}_i}$'s are mutually disjoint, and each one is disjoint from the alphabet $\mathcal{A}_\mathcal{G}$. We assume that each local ontology is expressed as a theory (named simply $\mathcal{S}_i$) in some logic $\mathcal{L}_{S_i}$, and we use $\mathcal{S}$ to denote the collection of theories $\mathcal{S}_1, \ldots, \mathcal{S}_n$.

Mapping. The mapping $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ is the heart of the OIS, in that it specifies how the concepts[1] in the global ontology $\mathcal{G}$ and in the local ontologies $\mathcal{S}$ map to each other.

Semantics. Intuitively, in specifying the semantics of an OIS, we have to start with a model of the local ontologies, and the crucial point is to specify which are the models of the global ontology. Thus, for assigning semantics to an OIS $\mathcal{O} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M}_{\mathcal{G},\mathcal{S}} \rangle$, we start by considering a *local model* $\mathcal{D}$ for $\mathcal{O}$, i.e., an interpretation that is a model for all the theories of $\mathcal{S}$. We call *global interpretation* for $\mathcal{O}$ any interpretation for $\mathcal{G}$. A global interpretation $\mathcal{I}$ for $\mathcal{O}$ is said to be a *global model for $\mathcal{O}$ wrt $\mathcal{D}$* if:

- $\mathcal{I}$ is a model of $\mathcal{G}$, and
- $\mathcal{I}$ satisfies the mapping $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$.

In the next sections, we will come back to the notion of satisfying a mapping wrt a local model. The semantics of $\mathcal{O}$, denoted $sem(\mathcal{O})$, is defined as follows:

$$sem(\mathcal{O}) = \{\, \mathcal{I} \mid \text{there exists a local model } \mathcal{D} \text{ for } \mathcal{O} \text{ s.t. } \mathcal{I} \text{ is a global model for } \mathcal{O} \text{ wrt } \mathcal{D} \,\}$$

Queries. Queries posed to an OIS $\mathcal{O}$ are expressed in terms of a query language $\mathcal{Q}_{\mathcal{G}}$ over the alphabet $\mathcal{A}_{\mathcal{G}}$ and are intended to extract a set of tuples of elements of Δ. Thus, every query has an associated arity, and the semantics of a query q of arity n is defined as follows. The answer $q^{\mathcal{O}}$ of q to $\mathcal{O}$ is the set of tuples

$$q^{\mathcal{O}} = \{\langle c_1, \ldots, c_n \rangle \mid \text{for all } \mathcal{I} \in sem(\mathcal{O}),\ \langle c_1, \ldots, c_n \rangle \in q^{\mathcal{I}} \,\}$$

where $q^{\mathcal{I}}$ denotes the result of evaluating q in the interpretation $\mathcal{I}$.

As we said before, the mapping $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ represents the heart of an OIS $\mathcal{O} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M}_{\mathcal{G},\mathcal{S}} \rangle$. In the usual approaches to ontology integration, the mechanisms for specifying the mapping between concepts in different ontologies are limited to expressing direct correspondences between terms. We argue that, in a real-world setting, one needs a much more powerful mechanism. In particular, such a mechanism should allow for mapping a concept in one ontology into a *view*, i.e., a query over the other ontologies, which acquires the relevant information by navigating and aggregating several concepts.

Following the research done in data integration [16, 17], we can distinguish two basic approaches for defining this mapping:

- the *global-centric approach*, where concepts of the global ontology $\mathcal{G}$ are mapped into queries over the local ontologies in $\mathcal{S}$;
- the *local-centric approach*, where concepts of the local ontologies in $\mathcal{S}$ are mapped to queries over the global ontology $\mathcal{G}$.

We discuss these two approaches in the following sections.

[1] Here and below we use the term "concept" for denoting a concept of the ontology.

3 Global-centric approach

In the global-centric approach (aka global-as-view approach), we assume we have a query language $\mathcal{V}_\mathcal{S}$ over the alphabet $\mathcal{A}_\mathcal{S}$, and the mapping between the global and the local ontologies is given by associating to each term in the global ontology a *view*, i.e., a query, over the local ontologies. The intended meaning of associating to a term C in $\mathcal{G}$ a query V_s over $\mathcal{S}$, is that such a query represents the best way to characterize the instances of C using the concepts in $\mathcal{S}$. A further mechanism is used to specify if the correspondence between C and the associated view is *sound*, *complete*, or *exact*. Let $\mathcal{D}$ be a local model for $\mathcal{O}$, and $\mathcal{I}$ a global interpretation for $\mathcal{O}$:

- $\mathcal{I}$ satisfies the correspondence $\langle C, V_s, sound\rangle$ in $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$, if all the tuples satisfying V_s in $\mathcal{D}$ satisfy C in $\mathcal{I}$,
- $\mathcal{I}$ satisfies the correspondence $\langle C, V_s, complete\rangle$ in $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$, if no tuple other than those satisfying V_s in $\mathcal{D}$ satisfies C in $\mathcal{I}$.
- $\mathcal{I}$ satisfies the correspondence $\langle C, V_s, exact\rangle$ in $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$, if the set of tuples that satisfy C in $\mathcal{I}$ is exactly the set of tuples satisfying V_s in $\mathcal{D}$.

We say that $\mathcal{I}$ *satisfies* the mapping $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$, if $\mathcal{I}$ satisfies every correspondence in $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$.

The global-centric approach is the one adopted in most data integration systems. In such systems, sources are databases (in general relational ones), the global ontology is actually a database schema (again, represented in relational form), and the mapping is specified by associating to each relation in the global schema one relational query over the source relations. It is a common opinion that this mechanism allow for a simple query processing strategy, which basically reduces to unfolding the query using the definition specified in the mapping, so as to translate the query in terms of accesses to the sources [20]. Actually, when we add constraints (even of a very simple form) to the global schema, query processing becomes even harder, as shown in the following case study.

3.1 A case study

We now set up a global-centric framework for ontology integration, which is based on ideas developed for data integration over global schemas expressed in the Entity-Relationship model [3]. In particular, we describe the main components of the ontology integration system, and we provide the semantics both of the system, and of query answering.

The OIS $O = \langle \mathcal{G}, \mathcal{S}, \mathcal{M}_{\mathcal{G},\mathcal{S}}\rangle$ is defined as follows:

- The *global ontology* $\mathcal{G}$ is expressed in the *Entity-Relationship model* (or equivalently as *UML class diagrams*). In particular, $\mathcal{G}$ may include:
  - typing constraints on relationships, assigning an entity to each component of the relationship;
  - mandatory participation to relationships, saying that each instance of an entity must participate as i-th component to a relationship;
  - ISA relations between both entities and relationships;

- typing constraints, functional restrictions, and mandatory existence, for attributes both of entities and of relationships.

- The *local ontologies* $\mathcal{S}$ are constituted simply by a relational alphabet $\mathcal{A}_{\mathcal{S}}$, and by the extensions of the relations in $\mathcal{A}_{\mathcal{S}}$. For example, such extensions may be expressed as relational databases. Observe that we are assuming that no intensional relation between terms in $\mathcal{A}_{\mathcal{S}}$ is present in the local ontologies.

- The *mapping* $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ between $\mathcal{G}$ and $\mathcal{S}$ is given by a set of correspondences of the form $\langle C, V_s, sound\rangle$, where C is a concept (i.e., either an entity, a relationship, or an attribute) in the global ontology and V_s is a query over $\mathcal{S}$. More precisely,

  - The mapping associates a query of arity 1 to each entity of $\mathcal{G}$.
  - The mapping associates a query of arity 2 to each entity attribute A of $\mathcal{G}$. Intuitively, if the query retrieves the pair $\langle x, y\rangle$ from the extension of the local ontologies, this means that y is a value of the attribute A of the entity instance x. Thus, the first argument of the query corresponds to the instances of the entity for which A is defined, and the second argument corresponds to the values of the attribute A.
  - The mapping associates a query of arity n to each relationship R of arity n in $\mathcal{G}$. Intuitively, if the query retrieves the tuple $\langle x_1, \ldots, x_n\rangle$ from the extension of the local ontologies, this means that $\langle x_1, \ldots, x_n\rangle$ is an instance of R.
  - The mapping associates a query of arity $n + 1$ to each attribute A of a relationship R of arity n in $\mathcal{G}$. The first n arguments of the query correspond to the tuples of R, and the last argument corresponds to the values of A.

  As specified above, the intended meaning of the query $\mathcal{V}_s$ associated to the concept C is that it specifies how to retrieve the data corresponding to C in the global schema starting from the data at the sources. This confirms that we are following the global-as-views approach: each concept in the global ontology is defined as a view over the concepts in the local ontologies. We do not pose any constraint on the language used to express the queries in the mapping. Since the extensions of local ontologies are relational databases, we simply assume that the language is able to express computations over relational databases.

To specify the semantics of a data integration system, we have to characterize, given the set of tuples in the extension of the various relations of the local ontologies, which are the data satisfying the global ontology. In principle, one would like to have a single extension as model of the global ontology. Indeed, this is the case for most of the data integration systems described in the literature. However, we will show in the following the surprising result that, due to the presence of the semantic conditions that are implicit in the conceptual schema $\mathcal{G}$, in general, we will have to account for a set of possible extensions.

Example 1. Figure 1 shows the global schema $\mathcal{G}_1$ of a data integration system $\mathcal{O}_1 = \langle\mathcal{G}_1, \mathcal{S}_1, \mathcal{M}_1\rangle$, where Age is a functional attribute, Student has a mandatory participation in the relationship Enrolled, Enrolled isa Member, and University isa Organization. The schema models persons who can be members of one or more organizations, and students who are

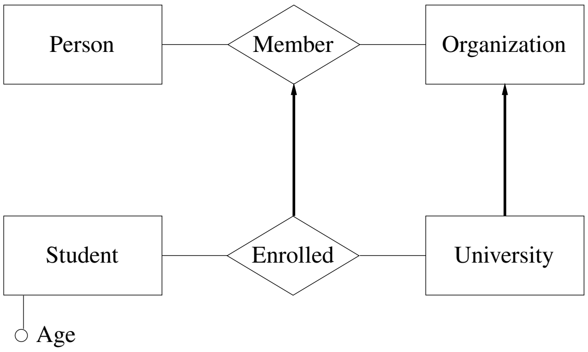

Figure 1: Global ontology of Example 1

enrolled in universities. Suppose that $\mathcal{S}$ is constituted by $\mathsf{S}_1, \mathsf{S}_2, \mathsf{S}_3, \mathsf{S}_4, \mathsf{S}_5, \mathsf{S}_6, \mathsf{S}_7, \mathsf{S}_8$, and that the mapping $\mathcal{M}_1$ is as follows:

$$
\begin{array}{rcl}
\mathsf{Person}(x) & \leftarrow & \mathsf{S}_1(x) \\
\mathsf{Organization}(x) & \leftarrow & \mathsf{S}_2(x) \\
\mathsf{Member}(x, y) & \leftarrow & \mathsf{S}_7(x, z) \wedge \mathsf{S}_8(z, y) \\
\mathsf{Student}(x) & \leftarrow & \mathsf{S}_3(x, y) \vee \mathsf{S}_4(x) \\
\mathsf{Age}(x, y) & \leftarrow & \mathsf{S}_3(x, y) \vee \mathsf{S}_6(x, y, z) \\
\mathsf{University}(x) & \leftarrow & \mathsf{S}_5(x) \\
\mathsf{Enrolled}(x, y) & \leftarrow & \mathsf{S}_4(x, y)
\end{array}
$$

■

From the semantics of the OIS $\mathcal{O}$ it is easy to see that, given a local model $\mathcal{D}$, several situations are possible:

1. No global model exists. This happens, in particular, when the data in the extension of the local ontologies retrieved by the queries associated to the elements of the global ontology do not satisfy the functional attribute constraints.

2. Several global models exist. This happens, for example, when the data in the extension of the local ontologies retrieved by the queries associated to the global concepts do not satisfy the ISA relationships of the global ontology. In this case, it may happen that several ways exist to add suitable objects to the elements of $\mathcal{G}$ in order to satisfy the constraints. Each such ways yields a global model.

Example 2. Referring to Example 1, consider a local model $\mathcal{D}_1$, where S_3 contains the tuple $\langle \mathsf{t}_1, \mathsf{a}_1 \rangle$, and S_6 contains the tuple $\langle \mathsf{t}_1, \mathsf{a}_2, \mathsf{v}_1 \rangle$. The query associated to Age by the mapping $\mathcal{M}_1$ specifies that, in every model of $\mathcal{O}_1$ both tuples should belong to the extension of Age. However, since Age is a functional attribute in $\mathcal{G}_1$, it follows that no model exists for the OIS $\mathcal{O}_1$. ■

Example 3. Referring again to Example 1, consider a local model $\mathcal{D}_2$, where S_1 contains p_1 and p_2, S_2 contains o_1, S_5 contains u_1, S_4 contains t_1, and the pairs $\langle \mathsf{p}_1, \mathsf{o}_1 \rangle$ and $\langle \mathsf{p}_2, \mathsf{u}_1 \rangle$ are in the join between S_7 and S_8. By the mapping $\mathcal{M}_1$, it follows that in every model of $\mathcal{O}_1$, we have that $\mathsf{p}_1, \mathsf{p}_2 \in \mathsf{Person}$, $\langle \mathsf{p}_1, \mathsf{o}_1 \rangle, \langle \mathsf{p}_2, \mathsf{u}_1 \rangle \in \mathsf{Member}$, $\mathsf{o}_1 \in \mathsf{Organization}$, $\mathsf{t}_1 \in \mathsf{Student}$, and $\mathsf{u}_1 \in \mathsf{University}$. Moreover, since $\mathcal{G}_1$ specifies that Student has a mandatory participation in the relationship Enrolled, in every model for $\mathcal{O}_1$, t_1 *must* be enrolled in a certain university. The key point is that nothing is said in $\mathcal{D}_2$ about *which* university, and therefore we have to accept as models all interpretations for $\mathcal{O}_1$ that differ in the university t_1 is enrolled in. ■

In the framework proposed, it is assumed that the first problem is solved by the queries extracting data from the extension of the local ontologies. In other words, it is assumed that, for any functional attribute A, the corresponding query implements a suitable data cleaning strategy (see, e.g., [15]) that ensures that, for every local model $\mathcal{D}$ and every x, there is at most one tuple (x, y) in the extension of A (a similar condition holds for functional attributes of relationships).

The second problem shows that the issue of query answering with incomplete information arises even in the global-as-view approach to data integration. Indeed, the existence of multiple global models for the OIS implies that query processing cannot simply reduce to evaluating the query over a single relational database. Rather, we should in principle take *all* possible global models into account when answering a query.

It is interesting to observe that there are at least two different strategies to simplify the setting, and overcome this problem that are frequently adopted in data integration systems [16, 20, 17]:

- Data integration systems usually adopt a simpler data model (often, a plain relational data model) for expressing the global schema (i.e., the global ontology). In this case, the data retrieved from the sources (i.e., the local ontologies) trivially fits into the schema, and can be directly considered as the unique database to be processed during query answering.

- The queries associated to the concepts of the global schema are often considered as exact. In this case, analogously to the previous one, it is easy to see that the only global extension to be considered is the one formed by the data retrieved by the extension of the local ontologies. However, observe that, when data in this extension do not obey all semantic conditions that are implicit in the global ontology, this single extension is not coherent with the global ontology, and the OIS is inconsistent. This implies that query answering in meaningless. We argue that, in the usual case of autonomous, heterogeneous local ontologies, it is very unlikely that data fit in the global ontology, and therefore, this approach is too restrictive, in the sense that the OIS would be often inconsistent.

The fact that the problem of incomplete information is overlooked in current approaches can be explained by observing that traditional data integration systems follow one of the above mentioned simplifying strategies: they either express the global schema as a set of plain relations, or consider the sources as exact (see, for instance, [11, 19, 1]).

In [3] we present an algorithm for computing the set of certain answers to queries posed to a data integration system. The key feature of the algorithm is to reason about both the query and the global ontology in order to infer which tuples satisfy the query in all models of the OIS. Thus, the algorithm does not simply unfold the query on the basis of the mapping, as usually done in data integration systems based on the global-as-view approach. Indeed, the algorithm is able to add more answers to those directly extracted from the local ontologies, by exploiting the semantic conditions expressed in the conceptual global schema.

Let $\mathcal{O} = \langle \mathcal{G}, \mathcal{S}, \mathcal{M}_{\mathcal{G},\mathcal{S}} \rangle$ be an OIS, let $\mathcal{D}$ be a local model, and let Q be a query over the global ontology $\mathcal{G}$. The algorithm is constituted by three major steps.

1. From the query Q, obtain a new query $expand_{\mathcal{G}}(Q)$ over the elements of the global ontology G in which the knowledge in $\mathcal{G}$ that is relevant for Q has been compiled in.

2. From $expand_{\mathcal{G}}(Q)$, compute the query $unfold_{\mathcal{M}_{\mathcal{G},\mathcal{S}}}(expand_{\mathcal{G}}(Q))$, by unfolding $expand_{\mathcal{G}}(Q)$ on the basis of the mapping $\mathcal{M}_{\mathcal{G},\mathcal{S}}$. The unfolding simply substitutes each atom of $expand_{\mathcal{G}}(Q)$ with the query associated by $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ to the element in the atom. The resulting $unfold_{\mathcal{M}_{\mathcal{G},\mathcal{S}}}(expand_{\mathcal{G}}(Q))$ is a query over the relations in the local ontologies.

3. Evaluate the query $unfold_{\mathcal{M}_{\mathcal{G},\mathcal{S}}}(expand_{\mathcal{G}}(Q))$ over the local model $\mathcal{D}$.

The last two steps are quite obvious. Instead, the first one requires to find a way to compile into the query the semantic relations holding among the concepts of the global schema $\mathcal{G}$. A way to do so is shown in [3]. The query $expand_{\mathcal{G}}(Q)$ returned by the algorithm is exponential wrt to Q. However, $expand_{\mathcal{G}}(Q)$ is a union of conjunctive queries, which, if the queries in the mapping are polynomial, makes the entire algorithm polynomial in data complexity.

Example 4. Referring to Example 3, consider the query Q_1 to $\mathcal{O}_1$:

$$\mathsf{Q}_1(x) \leftarrow \mathsf{Member}(x, y) \wedge \mathsf{University}(y)$$

It is easy to see that $\{\mathsf{p}_2, \mathsf{t}_1\}$ is the set of certain answers to Q_1 with respect to $\mathcal{O}_1$ and $\mathcal{D}_2$. Thus, although $\mathcal{D}_2$ does not indicate in which university t_1 is enrolled, the semantics of $\mathcal{O}_1$ specifies that t_1 is enrolled in *a* university in all legal database for $\mathcal{O}_1$. Since Member is a generalization of Enrolled, this implies that t_1 is in $\mathsf{Q}_1^{\mathcal{O}}$, and hence is in $unf_{\mathcal{M}_1}(exp_{\mathcal{G}_1}(\mathsf{Q}_1))$ evaluated over $\mathcal{D}_2$. ■

4 Local-centric approach

In the local-centric approach (aka local-as-view approach), we assume we have a query language $\mathcal{V}_{\mathcal{G}}$ over the alphabet $\mathcal{A}_{\mathcal{G}}$, and the mapping between the global and the local ontologies is given by associating to each term in the local ontologies a *view*, i.e., a query over the global ontology. Again, the intended meaning of associating to a term C in $\mathcal{S}$ a query V_g over $\mathcal{G}$, is that such a query represents the best way to characterize the instances of C using the concepts in $\mathcal{G}$. As in the global-centric approach, the correspondence between C and the associated view can be either sound, complete, or exact. Let $\mathcal{D}$ be a local model for $\mathcal{O}$, and $\mathcal{I}$ a global interpretation for $\mathcal{O}$:

- $\mathcal{I}$ satisfies the correspondence $\langle V_g, C, sound\rangle$ in $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$, if all the tuples satisfying C in $\mathcal{D}$ satisfy V_g in $\mathcal{I}$,

- $\mathcal{I}$ satisfies the correspondence $\langle V_g, C, complete\rangle$ in $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$, if no tuple other than those satisfying C in $\mathcal{D}$ satisfies V_g in $\mathcal{I}$,

- $\mathcal{I}$ satisfies the correspondence $\langle V_g, C, exact\rangle$ in $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$, if the set of tuples that satisfy C in $\mathcal{D}$ is exactly the set of tuples satisfying V_g in $\mathcal{I}$.

As in the global-centric approach, we say that $\mathcal{I}$ *satisfies* the mapping $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$, if $\mathcal{I}$ satisfies every correspondence in $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$.

Recent research work on data integration follows the local-centric approach [20, 17, 18, 6, 8]. The major challenge of this approach is that, in order to answer a query expressed over the global schema, one must be able to reformulate the query in terms of queries to the sources. While in the global-centric approach such a reformulation is guided by the correspondences in the mapping, here the problem requires a reasoning step, so as to infer how to use the sources for answering the query. Many authors point out that, despite its difficulty, the local-centric approach better supports a dynamic environment, where local ontologies can be added to the systems without the need for restructuring the global ontology.

4.1 A case study

We present here an OIS architecture based on the use of Description Logics to represent ontologies [6, 7]. Specifically, we adopt the Description Logic $\mathcal{DLR}$, in which both classes and n-ary relations can be represented [4]. We first introduce $\mathcal{DLR}$, and then we illustrate how we use the logic to define an OIS.

4.1.1 The Description Logic $\mathcal{DLR}$

Description Logics[2] (DLs) are knowledge representation formalisms that are able to capture virtually all class-based representation formalisms used in Artificial Intelligence, Software Engineering, and Databases [9, 10].

One of the distinguishing features of these logics is that they have optimal reasoning algorithms, and practical systems implementing such algorithms are now used in several projects.

In DLs, the domain of interest is modeled by means of *concepts* and *relations*, which denote classes of objects and relationships, respectively. Here, we focus our attention on the DL $\mathcal{DLR}$ [4, 6], whose basic elements are *concepts* (unary relations), and *n-ary relations*. We assume to deal with an alphabet $\mathcal{A}$ constituted by a finite set of atomic relations, atomic concepts, and *constants*, denoted by P, A, and a, respectively. We use R to denote arbitrary relations (of given arity between 2 and n_{max}), and C to denote arbitrary concepts, respectively built according to the following syntax:

$$\begin{array}{lcl} C & ::= & \top_1 \mid A \mid \neg C \mid C_1 \sqcap C_2 \mid \exists[i]R \mid (\leq k\,[i]R) \\ R & ::= & \top_n \mid P \mid i/n\!:\!C \mid \neg R \mid R_1 \sqcap R_2 \end{array}$$

where i denotes a component of a relation, i.e., an integer between 1 and n_{max}, n denotes the *arity* of a relation, i.e., an integer between 2 and n_{max}, and k denotes a nonnegative integer. We consider only concepts and relations that are *well-typed*, which means that only relations of the same arity n are combined to form expressions of type $R_1 \sqcap R_2$ (which inherit the arity n), and $i \leq n$ whenever i denotes a component of a relation of arity n.

The semantics of $\mathcal{DLR}$ is specified as follows. An *interpretation* $\mathcal{I}$ is constituted by an *interpretation domain* $\Delta^{\mathcal{I}}$, and an *interpretation function* $\cdot^{\mathcal{I}}$ that assigns to each constant an element of $\Delta^{\mathcal{I}}$ under the unique name assumption, to each concept C a subset $C^{\mathcal{I}}$ of $\Delta^{\mathcal{I}}$, and to each relation R of arity n a subset $R^{\mathcal{I}}$ of $(\Delta^{\mathcal{I}})^n$, such that the conditions in Figure 2 are satisfied. Observe that, the "$\neg$" constructor on relations is used to express difference of relations, and not the complement [4].

A $\mathcal{DLR}$ knowledge base is a set of inclusion assertions of the form

$$C_1 \sqsubseteq C_2 \qquad\qquad R_1 \sqsubseteq R_2$$

[2]See `http://dl.kr.org` for the home page of Description Logics.

$$
\begin{array}{rcl}
\top_1^{\mathcal{I}} & = & \Delta^{\mathcal{I}} \\
A^{\mathcal{I}} & \subseteq & \Delta^{\mathcal{I}} \\
(\neg C)^{\mathcal{I}} & = & \Delta^{\mathcal{I}} \setminus C^{\mathcal{I}} \\
(C_1 \sqcap C_2)^{\mathcal{I}} & = & C_1^{\mathcal{I}} \cap C_2^{\mathcal{I}} \\
(\exists[i]R)^{\mathcal{I}} & = & \{d \in \Delta^{\mathcal{I}} \mid \exists \langle d_1, \ldots, d_n \rangle \in R^{\mathcal{I}}. d_i = d\} \\
(\leq k\,[i]R)^{\mathcal{I}} & = & \{d \in \Delta^{\mathcal{I}} \mid \sharp\{\langle d_1, \ldots, d_n \rangle \in R_1^{\mathcal{I}} \mid d_i = d\} \leq k\} \\
\top_n^{\mathcal{I}} & \subseteq & (\Delta^{\mathcal{I}})^n \\
P^{\mathcal{I}} & \subseteq & \top_n^{\mathcal{I}} \\
i/n : C^{\mathcal{I}} & = & \{\langle d_1, \ldots, d_n \rangle \in \top_n^{\mathcal{I}} \mid d_i \in C^{\mathcal{I}}\} \\
(\neg R)^{\mathcal{I}} & = & \top_n^{\mathcal{I}} \setminus R^{\mathcal{I}} \\
(R_1 \sqcap R_2)^{\mathcal{I}} & = & R_1^{\mathcal{I}} \cap R_2^{\mathcal{I}}
\end{array}
$$

Figure 2: Semantic rules for $\mathcal{DLR}$ (P, R, R_1, and R_2 have arity n)

where C_1 and C_2 are concepts, and R_1 and R_2 are relations of the same arity. An inclusion assertion $C_1 \sqsubseteq C_2$ (resp., $R_1 \sqsubseteq R_2$) is satisfied in an interpretation $\mathcal{I}$ if $C_1^{\mathcal{I}} \subseteq C_2^{\mathcal{I}}$ (resp., $R_1^{\mathcal{I}} \subseteq R_2^{\mathcal{I}}$). An interpretation is a *model* of a knowledge base $\mathcal{K}$, if it satisfies all assertions in $\mathcal{K}$. $\mathcal{K}$ *logically implies* an inclusion assertion ρ if ρ is satisfied in all models of $\mathcal{K}$.

Finally, we introduce the notion of query expression in $\mathcal{DLR}$. We assume that the alphabet $\mathcal{A}$ is enriched with a finite set of variable symbols, simply called *variables*. A *query expression* Q over a $\mathcal{DLR}$ knowledge base $\mathcal{K}$ is a non-recursive datalog query of the form

$$Q(\vec{x}) \leftarrow conj_1(\vec{x}, \vec{y}_1) \vee \cdots \vee conj_m(\vec{x}, \vec{y}_m)$$

where each $conj_i(\vec{x}, \vec{y}_i)$ is a conjunction of *atoms*, and $\vec{x}$, $\vec{y}_i$ are all the variables appearing in the conjunct. Each atom has one of the forms $R(\vec{t})$ or $C(t)$, where $\vec{t}$ and t are variables in $\vec{x}$ and $\vec{y}_i$ or constants in $\mathcal{A}$, R is a relation of $\mathcal{K}$, and C is a concept of $\mathcal{K}$. The number of variables of $\vec{x}$ is called the *arity* of Q, and is the arity of the relation denoted by the query Q. We observe that the atoms in query expressions are arbitrary $\mathcal{DLR}$ concepts and relations, freely used in the assertions of the KB.

Given an interpretation $\mathcal{I}$, a query expression Q of arity n is interpreted as the set $Q^{\mathcal{I}}$ of n-tuples of constants $\langle c_1, \ldots, c_n \rangle$, such that, when substituting each c_i for x_i, the formula

$$\exists \vec{y}_1 . conj_1(\vec{x}, \vec{y}_1) \vee \cdots \vee \exists \vec{y}_m . conj_m(\vec{x}, \vec{y}_m)$$

evaluates to true in $\mathcal{I}$.

$\mathcal{DLR}$ is equipped with effective reasoning techniques that are sound and complete with respect to the semantics. In particular, checking whether a given assertion logically follows from a set of assertions is EXPTIME-complete in (assuming that numbers are encoded in unary), and query containment, i.e., checking whether one query is contained in another one in every model of a set of assertions, is EXPTIME-hard and solvable in 2EXPTIME [4].

4.1.2 $\mathcal{DLR}$ local-centric OIS

We now set up a local-centric framework for ontology integration, which is based on ideas developed for data integration over $\mathcal{DLR}$ knowledge bases [6, 5]. In particular, we describe the main components of the ontology integration system, and we provide the semantics both of the system, and of query answering.

In this setting, an OIS $O = \langle \mathcal{G}, \mathcal{S}, \mathcal{M}_{\mathcal{G},\mathcal{S}} \rangle$ is defined as follows:

- The *global ontology* $\mathcal{G}$ is a $\mathcal{DLR}$ knowledge base.
- The *local ontologies* $\mathcal{S}$ are again seen as a set of relations each giving the extension of an ontology-concept in the ontology. We observe that again we have only extensional knowledge on such relations in $\mathcal{S}$.
- The *mapping* $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ between $\mathcal{G}$ and $\mathcal{S}$ is given by a set of correspondences of the form $\langle V_g, T, as \rangle$, where T is a relation of a local ontology, $\mathcal{V}_g$ is a query expression over $\mathcal{G}$, and as is either $sound$, $complete$, or $exact$.

Observe that we could partition the global ontology in several parts, one for each local ontology, modeling the intensional knowledge on the local ontology wrt the OIS, plus one for the reconciled global view of such ontologies. By making use of the so called interschema assertions [12] the different parts can be related to each at the intesional level. For simplicity we do not deal with interschema assertion in this case study, however it is immediate to extend the framework presented here to include them as well [6, 7].

Query answering in this setting requires quite sophisticated techniques that take into account the knowledge both in the global ontology and in the mapping in answering a query posed over the global ontology with the data contained in the local ontologies. Such query answering techniques are studied in [5].

Example 5. Consider for example the OIS $\mathcal{O}_d = \langle \mathcal{G}_d, \mathcal{S}_d, \mathcal{M}_d \rangle$ defined as follows:

- The global ontology $\mathcal{G}_d$ is the $\mathcal{DLR}$ knowledge base

  $$\begin{array}{rcl} \mathsf{American} \sqcap \exists[1](\mathsf{RELATIVE} \sqcap 2 : \mathsf{Doctor}) & \sqsubseteq & \mathsf{Wealthy} \\ \mathsf{Surgeon} & \sqsubseteq & \mathsf{Doctor} \end{array}$$

  expressing that Americans who have a doctor as relative are wealthy, and that each surgeon is also a doctor.

- The set $\mathcal{S}_d$ of local ontologies consists of two ontologies, containing respectively the relations T_1 and T_2, with extensions $\{\mathsf{ann}, \mathsf{bill}\}$ and $\{\mathsf{ann}, \mathsf{dan}\}$.
- The mapping $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ is $\{\langle \mathsf{V}_1, \mathsf{T}_1, sound \rangle,\ \langle \mathsf{V}_2, \mathsf{T}_2, sound \rangle\}$, with

  $$\begin{array}{rcl} \mathsf{V}_1(x) & \leftarrow & \mathsf{RELATIVE}(x, y) \wedge \mathsf{Surgeon}(y) \\ \mathsf{V}_2(x) & \leftarrow & \mathsf{American}(x) \end{array}$$

Given the query expression $\mathsf{Q}_w(x) \leftarrow \mathsf{Wealthy}(x)$ over $\mathcal{G}_d$, asking for those who are wealthy, we have that the only answer in $\mathsf{Q}_w^{\mathcal{O}_d}$ is ann. Consider an additional local ontology, consisting of a relation T_3 with an extension not containing bill, and mapped to $\mathcal{G}$ by the correspondence $\langle \mathsf{V}_3, \mathsf{T}_3, exact \rangle$, with $\mathsf{V}_3(x) \leftarrow \mathsf{Wealthy}(x)$. Then, from the constraints in $\mathcal{G}_d$ and the information we have on the correspondences, we can conclude that bill is not an answer to the query asking for the Americans. ∎

5 Combining the global-centric and local-centric approaches

The global-centric and the local-centric approach can be combined together into an approach using unrestricted mappings, in which the restrictions on the direction of the correspondence between global and local ontologies are overcome [14]. In the unrestricted approach, we have both a query language $\mathcal{V}_{\mathcal{S}}$ over the alphabet $\mathcal{A}_{\mathcal{S}}$, and a query language $\mathcal{V}_{\mathcal{G}}$ over the alphabet $\mathcal{A}_{\mathcal{G}}$, and the mapping between the global and the local ontologies is given by relating views over the global ontology to views over the local ontologies. Again, the intended meaning of relating the view V_g over the global ontology to the view V_s over the local ontology is that V_s represents the best way to characterize the objects satisfying V_g in terms of the concepts in $\mathcal{S}$. Analogously to the other cases, the correspondences between V_g and V_s can be characterized as sound, complete, or exact. Let $\mathcal{D}$ be a local model for $\mathcal{O}$, and $\mathcal{I}$ a global interpretation for $\mathcal{O}$:

- $\mathcal{I}$ satisfies the correspondence $\langle V_g, V_s, sound\rangle$ in $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$, if all the tuples satisfying satisfying V_s in $\mathcal{D}$ satisfy V_g in $\mathcal{I}$,
- $\mathcal{I}$ satisfies the correspondence $\langle V_g, V_s, complete\rangle$ in $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$, if no tuple other than those satisfying V_s in $\mathcal{D}$ satisfy V_g in $\mathcal{I}$,
- $\mathcal{I}$ satisfies the correspondence $\langle V_g, V_s, exact\rangle$ in $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$, if the set of tuples that satisfy V_g in $\mathcal{I}$ is exactly the set of tuples satisfying V_s in $\mathcal{D}$.

Again, we say that $\mathcal{I}$ *satisfies* the mapping $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$, if $\mathcal{I}$ satisfies every correspondence in $\mathcal{M}_{\mathcal{G},\mathcal{S}}$ wrt $\mathcal{D}$.

Example 6. Consider the OIS $\mathcal{O}_u = \langle \mathcal{G}_u, \mathcal{S}_u, \mathcal{M}_u\rangle$, where both $\mathcal{G}_u$ and the two ontologies S_1 and S_2 forming $\mathcal{S}_u$ are simply sets of relations with their extensions.

- The global ontology $\mathcal{G}_u$ contains two binary relations, WorksFor, denoting researchers and projects they work for, and Area, denoting projects and research areas they belong to.
- The local ontology S_1 contains a binary relation InterestedIn denoting persons and fields they are interested in, and the local ontology S_2 contains a binary relation GetGrant, denoting researchers and grants assigned to them, and a binary relation GrantFor denoting grants and projects they refer to.
- The mapping $\mathcal{M}_u$ is formed by the following correspondences
  - $\langle \mathsf{V}_1, \mathsf{InterestedIn}, complete\rangle$, with $\mathsf{V}_1(r, f) \leftarrow \mathsf{WorksFor}(r, p) \wedge \mathsf{Area}(p, f)$
  - $\langle \mathsf{WorkFor}, \mathsf{V}_2, sound\rangle$, with $\mathsf{V}_2(r, p) \leftarrow \mathsf{GetGrant}(r, g) \wedge \mathsf{GrantFor}(g, p)$

This situation can be represented neither in the global-centric nor in the local-centric approach. ∎

Query answering in this approach is largely unexplored, mainly because it combines the difficulties of the other ones. However, in a real world setting, this may be the only approach that provides the appropriate expressive power.

6 Conclusions

We have presented a general framework for ontology integration, where a global ontology is used to provide a unified view for querying local ontologies, as in the semantic web. The framework represents a sort of design space for the problem of integrating ontologies within semantic web applications. We have argued that the mapping between the global and the local ontologies is the main aspect of the framework, and we have discussed various approaches for specifying such a mapping. Independently of the approach, we have stressed that the notion of query is crucial for the task of ontology integration.

The two case studies we have presented have shown the need of sophisticated techniques for query answering in an ontology integration system. The two case studies illustrated simplified settings, drawn from data integration. One should expect things to become even more complex when ontology integration is considered in its full generality. Recently several proposals have been made, based on the idea of expressing ontologies as knowledge bases, e.g., in Description Logics [13, 2], and applying automated reasoning techniques for several services in the design of and the interaction with the semantic web. We believe however that such an idea needs to be extended by considering queries as first order citizens and having the ability to reason on them.

References

[1] M. Bouzeghoub and M. Lenzerini. Special issue on data extraction, cleaning, and reconciliation. *Information Systems*, 2001. To appear.

[2] J. Broekstra, M. Klein, D. Fensel, and I. Horrocks. Adding formal semantics to the Web: building on top of RDF Schema. In *Proc. of the ECDL 2000 Workshop on the Semantic Web*, 2000.

[3] A. Calì, D. Calvanese, G. De Giacomo, and M. Lenzerini. Accessing data integration systems through conceptual schemas. In *Proc. of the 20th Int. Conf. on Conceptual Modeling (ER 2001)*, 2001. To appear.

[4] D. Calvanese, G. De Giacomo, and M. Lenzerini. On the decidability of query containment under constraints. In *Proc. of the 17th ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS'98)*, pages 149–158, 1998.

[5] D. Calvanese, G. De Giacomo, and M. Lenzerini. Answering queries using views over description logics knowledge bases. In *Proc. of the 17th Nat. Conf. on Artificial Intelligence (AAAI 2000)*, pages 386–391, 2000.

[6] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati. Description logic framework for information integration. In *Proc. of the 6th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'98)*, pages 2–13, 1998.

[7] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati. Information integration: Conceptual modeling and reasoning support. In *Proc. of the 6th Int. Conf. on Cooperative Information Systems (CoopIS'98)*, pages 280–291, 1998.

[8] D. Calvanese, G. De Giacomo, M. Lenzerini, and M. Y. Vardi. View-based query processing and constraint satisfaction. In *Proc. of the 15th IEEE Symp. on Logic in Computer Science (LICS 2000)*, pages 361–371, 2000.

[9] D. Calvanese, M. Lenzerini, and D. Nardi. Description logics for conceptual data modeling. In J. Chomicki and G. Saake, editors, *Logics for Databases and Information Systems*, pages 229–264. Kluwer Academic Publisher, 1998.

[10] D. Calvanese, M. Lenzerini, and D. Nardi. Unifying class-based representation formalisms. *J. of Artificial Intelligence Research*, 11:199–240, 1999.

[11] M. J. Carey, L. M. Haas, P. M. Schwarz, M. Arya, W. F. Cody, R. Fagin, M. Flickner, A. Luniewski, W. Niblack, D. Petkovic, J. Thomas, J. H. Williams, and E. L. Wimmers. Towards heterogeneous multimedia information systems: The Garlic approach. In *RIDE-DOM*, pages 124–131, 1995.

[12] T. Catarci and M. Lenzerini. Representing and using interschema knowledge in cooperative information systems. *J. of Intelligent and Cooperative Information Systems*, 2(4):375–398, 1993.

[13] S. Decker, D. Fensel, F. van Harmelen, I. Horrocks, S. Melnik, M. Klein, and J. Broekstra. Knowledge representation on the web. In *Proc. of the 2000 Description Logic Workshop (DL 2000)*, pages 89–97. CEUR Electronic Workshop Proceedings, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-33/, 2000.

[14] M. Friedman, A. Levy, and T. Millstein. Navigational plans for data integration. In *Proc. of the 16th Nat. Conf. on Artificial Intelligence (AAAI'99)*, pages 67–73. AAAI Press/The MIT Press, 1999.

[15] H. Galhardas, D. Florescu, D. Shasha, and E. Simon. An extensible framework for data cleaning. Technical Report 3742, INRIA, Rocquencourt, 1999.

[16] R. Hull. Managing semantic heterogeneity in databases: A theoretical perspective. In *Proc. of the 16th ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS'97)*, 1997.

[17] A. Y. Levy. Answering queries using views: A survey. Technical report, University of Washinghton, 1999.

[18] A. Y. Levy, D. Srivastava, and T. Kirk. Data model and query evaluation in global information systems. *J. of Intelligent Information Systems*, 5:121–143, 1995.

[19] C. Li, R. Yerneni, V. Vassalos, H. Garcia-Molina, Y. Papakonstantinou, J. D. Ullman, and M. Valiveti. Capability based mediation in TSIMMIS. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 564–566, 1998.

[20] J. D. Ullman. Information integration using logical views. In *Proc. of the 6th Int. Conf. on Database Theory (ICDT'97)*, volume 1186 of *Lecture Notes in Computer Science*, pages 19–40. Springer-Verlag, 1997.